Inauguraldissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften der Universität Mannheim

Testing econometric models with heterogeneous agents

Applications in treatment analysis and networks

Andreas Dzemski

Frühjahrssemester 2015

Abteilungssprecher: Prof. Dr. Eckhard Janeba
Referenten: Prof. Dr. Markus Frölich, Prof. Dr. Enno Mammen
Datum der mündlichen Prüfung: 7. Juli 2015

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbständig angefertigt und die benutzten Hilfsmittel vollständig und deutlich angegeben habe.

Mannheim, 8. Juni 2015

Andreas Dzemski

Curriculum Vitae

2010 - 2015	Ph.D. student at University of Mannheim
2008 - 2009	Visiting student at Yale University, New Haven
2005 - 2010	Undergraduate studies in Economics at University of Mannheim

Contents

Introduction

Economists use sophisticated empirical models to learn about agent behavior in the real world. Many such models are set up to account for unobserved agent characteristics that may drive agent behavior. In the econometric literature, such unobserved characteristics are referred to as *unobserved heterogeneity.* A crucial step in empirical research is model testing. To this end, the researcher has to find a way to check the predictions of her idealized model of agent behavior against the actions of agents observed in the real world. Such an endeavor is particularly difficult for models that allow for unobserved heterogeneity. To disprove such a model, the researcher has to make the case that the model cannot reproduce the observed data, regardless of which value the unobserved component takes.

This thesis tackles two concrete testing problems that carry substantial relevance in modern empirical research. In both cases, the presence of a flexible unobserved component invalidates popular testing procedures that have been developed for versions of the model that do not allow for unobserved heterogeneity or that make very restrictive assumptions on the nature of the unobserved heterogeneity. I suggest valid testing procedures and illustrate their empirical bite by applying them to real data. Moreover, I analyze a general framework for constructing model tests that is a applicable to a wide range of models that satisfy a latent index restriction. In this introduction I provide a non-technical overview of the research described in this thesis.

Chapter 1

In the first chapter (based on joint work with Florian Sarnetzki) I consider a testing problem from the literature of *treatment evaluation.* I suggest a procedure that can be used to test assumptions about so-called *instruments.* These are special observed variables that provide variation that can be used to estimate the causal effect of policy interventions.

The goal of treatment evaluation is to determine the benefit (or *treatment effect*) that an individual derives from exposure to a particular regime or treatment. Economically relevant treatments include a diverse set of conditions such as motherhood, obtaining an educational degree or unemployment. Treatment effects are unobserved, since each individual is observed in either its treated or its untreated state, but never in both states at the same time. I am studying the LATE-framework (Joshua Angrist, Imbens, and

Rubin 1996; Heckman and Vytlacil 2005; Vytlacil 2002) and assume that the assignment of individuals to the treatment and control groups is outside of the control of the researcher. Rather, selection into the treatment is *endogenous*, i.e., it is a part of the behavioral pattern that the model aims to predict. The model postulates that there is an unobserved component that drives participation decisions. Due to this component, two observationally equivalent agents may arrive at different participation decisions. The model places no restrictions on the relationship between the unobserved component in the participation decision and the treatment effect. For example, it permits for a setting in which the individuals with the largest treatment effects are also the most likely to be treated. Consequently, the treatment group and the control group need not be representative of the population. This is referred to as the *selection problem.* In particular, the observed difference in outcomes between the treatment and the control group will typically not give a valid estimator of the benefit of the treatment to the average individual.

To estimate treatment effects in this setting researchers use *instruments.* An instrument is an observed variable that satisfies two key assumptions. First, it has no effect on the treatment effect that the researcher is studying. Secondly, it changes agents' incentives to participate in the treatment in a way that is independent of unobserved heterogeneity. To illustrate this concept, consider one of the instruments discussed in this thesis. In Section 1.6, I consider the effect of motherhood on the probability of graduating from high-school. It seems plausible that unobserved factors such as family background may affect both the probability to enter motherhood (enter the treatment group) and the magnitude of the effect that motherhood has on the probability of graduating (the magnitude of the treatment effect). The variable indicating age at first menstrual period is a potential instrument. It randomly shifts female fecundity and it is unlikely to affect the likelihood of graduation (Ribar 1994). Clearly, by postulating that a variable is an instrument we are making a claim about how it interacts with the two components of unobserved heterogeneity, i.e., the unobserved treatment effects and the unobserved incentives to participate in the treatment. Testing such a claim is non-trivial. In particular, standard instrument tests, such as Hansen's overidentification test (L. P. Hansen 1982), that were developed for less convoluted selection problems do not apply (Heckman, Schmierer, and Urzua 2010).

The testing approach suggested in Chapter 1 exploits the fact that valid instruments affect the outcome only through the probability of participating in the treatment. This probability is known as the *propensity score.* To derive a testable restriction, suppose that two instruments are available. If both instruments are valid, then different instrument configurations that yield the same propensity score will also yield the same outcome. On the other hand, a variable that is falsely identified as an instrument may have an effect on outcomes that is not mediated through the propensity score. Therefore, an approach based on comparing observations with similar propensity scores and diverging instrument configurations can detect invalid instruments. Based on this testing idea, Chapter 1 presents a test statistic that is based on comparing predicted values between two sub-populations for a range of propensity scores. Precursors of this testing idea can be found in Heckman et al. 1998 and Vytlacil and Yildiz 2007.

The primary difficulty in implementing the test is that the propensity score of a given individual is not observed and has to be estimated. The additional estimation step affects the behavior of the test statistic and restricts the possible set-ups that endow the test with good theoretical properties. A substantial part of Chapter 1 is devoted to analyzing these restrictions. As a measure of the power of the test, I consider the rate at which it detects local alternatives. This measure is based on the notion that a good test should be able to catch small deviations from the null hypothesis.

Chapter 2

In the second chapter I generalize the testing problem from Chapter 1. The null hypothesis underlying the instrument test in Chapter 1 is that the propensity score aggregates all information provided by the instruments about the outcome. In the econometric literature, a function that maps a large vector into a lower dimensional space is often referred to as an *index*. The notion that all relevant information contained in the larger vector is accurately summarized by the index is known as *index sufficiency*. Thus, the instrument test furnishes an example of a test for index sufficiency with the propensity score serving as the index. As noted above, the main complication in implementing the instrument test is that the propensity score has to be estimated. The additional estimation step has to be accounted for when we judge the validity of the null hypothesis based on the observed value of the test statistic. We face the same problem in other testing situations where the null hypothesis postulates sufficiency of an index that is not observed but estimable.

In Chapter 2 I provide a test of index sufficiency with a predicted (estimated) index. The test statistic that I am considering is based on a test idea first developed in Delgado and Manteiga 2001. I adapt their test statistic to allow for a predicted index and I quantify how the additional estimation step changes the behavior of the test statistic in large samples.

The issue of accounting for estimation error from multiple stages is known as a generated regressors problem. Recently, the literature has devoted a substantial effort to advance the theory for estimation with generated regressors (Mammen, Rothe, and Schienle 2012; Mammen, Rothe, and Schienle 2015; Escanciano, Jacho-Chávez, and Lewbel 2014). Following an approach pioneered in recent research I do not assume a specific estimator with the ability to predict the index. Rather, my results apply to a varied class of estimators that satisfy certain restrictions.

The test statistic is carefully chosen to be well-behaved under a predicted index. In contrast to e.g. the test statistic considered in Chapter 2, it is based on unconditional moments rather than conditional moments. The additional integration step makes it easier to control the higher-order effects of estimating the index. In order to prove this advantage of my test statistic, I make extensive use of empirical process theory to develop new results for U-statistics with generated regressors.

As an illustration of possible gains from using the new framework, I investigate a testing problem that is very similar to the instrument test discussed in Chapter 1. In the new framework it is possible to quantify the impact of the additional estimation step without imposing strong assumptions about the bandwidth choices in two estimation stages. In particular, it is possible to use bandwidths that are optimal with respect to the mean integrated squared error (MISE). This is the bandwidth choice that is targeted by many data-driven methods for bandwidth selection.

Chapter 3

In the final chapter I consider agents that decide whether to enter a certain type of bilateral relationship with another agent. The bilateral relationship can be viewed as establishing a *directed link* between two agents. The collection of all such links in a prescribed community of agents is called the *network*. Agents are endowed with characteristics that determine their attractiveness as linking partners (*popularity*) and their willingness to link to other agents (*productivity*). These characteristics are observed by other agents in the network, who will take them into account when making their linking decisions. They are, however, typically not observed by the econometrician. In my linking model, accounting for this kind of unobserved heterogeneity poses a methodological challenge.

My model is a variation of the popular linking model by Holland and Leinhardt 1981. Unobserved agent heterogeneity is modeled by a fixed effects approach. This means that I do not require any assumptions about how popularity and productivity are distributed in the population and how they correlate with observed covariates. Agent popularity and productivity enter the estimation problem as additional parameters.

By observing the network it is possible to identify the unobserved popularity and productivity parameters. For example, a very popular agent will have more in-bound links than a less popular agent with similar observed characteristics. This way of estimating parameters deviates from standard estimation procedures. With every agent that is added another pair of parameters has to be estimated. Consequently, the informational value of adding another observation is different than it is in a standard set-up, where the number of estimated parameters stays constant as more observations are added. For the model that I am considering, the non-standard formulation of the estimation problem expresses itself as a so-called *incidental parameter problem* (Neyman and Scott 1948; Andersen 1970). This means that statistics calculated from the model are biased in a way that is not predicted by standard estimation theory. In the face of this problem, developing a test for the validity of the model is challenging.

The model test developed in Chapter 3 is based on the same idea as the test presented in Holland and Leinhardt 1978. It exploits the fact that the linking model predicts the likelihood of which link configurations occur within groups of several agents. Comparing the observed links within a group to the configuration predicted by the linking model is a way to refute the suggested model. In particular, my test considers the prevalence of a certain link configuration between three agents that is known as a *transitive triangle*. My specification test improves on previous tests (Holland and Leinhardt 1978; Karlberg 1997) in two ways. First, I allow for a substantial amount of unobserved heterogeneity in agents' linking decisions. Secondly, I explicitly account for the fact that the linking model (often referred to as reference distribution in the testing literature) is unknown and has to be estimated. In particular, I quantify the way in which the test statistic is affected by the incidental parameter problem and suggest a procedure for computing critical values. This procedure is based on a correction formula that can be applied to a naive version of the test statistic that disregards the presence of incidental parameters. The corrected test statistic follows a centered normal distribution.

In my model, agents make linking decisions based entirely on their own characteristics and on the characteristics of a potential linking partner. Such models are known as *dyadic models*. They have often been criticized due to their inability to account for the prevalence of certain link configuration within groups of agents. As a possible explanation, it has been suggested that agents act strategically in their linking decisions (Jackson and Wolinsky 1996) and actively work towards achieving certain configurations. I provide a stylized model that suggests that neglecting unobserved popularity effects in a dyadic model will lead a researcher to underestimate the prevalence of transitive triangles in a network. In studies that do not account for unobserved heterogeneity, the apparent abundance of transitive triangles has often led researchers to speculate that agents act strategically to satisfy a taste for transitive closure.

I apply my linking model and the corresponding specification test to network data from Indian villages. As expected, a model that does not account for unobserved heterogeneity does not predict a sufficient number of transitive triangles. However, once unobserved heterogeneity is added to the model, the test can no longer clearly distinguish the observed networks from the predicted networks. This suggests that unobserved heterogeneity might drive the emergence of certain network features that other models (Mele 2013; Leung 2014) attribute to strategic interaction.

Acknowledgements

I would like to express my gratitude to my two advisors Markus Frölich and Enno Mammen who did a terrific job in supporting me throughout the process and who always encouraged me to explore new ideas. My sincere thanks go to my co-author Florian Sarnetzki who helped me complete the first chapter of this thesis. I had the opportunity to present the research leading up to this thesis at various seminars, workshops and conferences and I am grateful for the numerous comments and suggestions that participants at theses events generously offered. Several individuals took the time to discuss research ideas and technical details with me or to provide detailed comments on my papers. In particular, I am grateful for comments from Peter Aronow, Allie Carnegie, Iván Fernández-Val, Bryan Graham, Martin Huber, Carsten Jentsch, Stephen Kastoryano, Patrick Kline, Jan Nimczik, Steffen Reinhold and Andrea Weber. I appreciate financial support from Landesgraduiertenförderung Baden-Württemberg and the Swiss-German research group FOR 916 (statistical regularization). Most of all, I would like to thank my wife Laura for keeping me sane, or at least as close to sane as you can hope for when taking on such a project.

CHAPTER 1

Overidentification test in a nonparametric treatment model with unobserved heterogeneity

with Florian Sarnetzki

1.1 Introduction

The canonical treatment effect evaluation problem in Economics can be phrased as the problem of recovering the coefficient β from the outcome equation

$$Y = \alpha + \beta D, \tag{1.1}$$

where D is a binary indicator of treatment status, and α and β are random coefficients. In latent outcome notation[1], the treatment effect β is commonly written as $\beta = Y^1 - Y^0$. If β is known to be constant, then it can be identified by classical instrumental variables methods. In this framework it is straightforward to test the validity of the instruments by classical GMM overidentification tests (L. P. Hansen 1982, Sargan 1958). In many applications the more natural assumption is to assume that the treatment effect β is non-constant and correlated with D. Economically this means that individuals differ in their gains from participating in the treatment and that when deciding whether to participate or not individuals take into account possible gains from participation. This setting is often referred to as one of *essential heterogeneity* (Heckman, Urzua, and Vytlacil 2006). It was first considered in the seminal papers by Imbens and Joshua Angrist 1994 and Joshua Angrist, Imbens, and Rubin 1996. These authors give assumptions under which a binary instrument identifies the average treatment effect for the subpopulation of compliers, which they dub the Local Average Treatment Effect (LATE). The compliers are the individuals that respond to a change in the realizations of the binary instrument by changing their participation decision. Different instruments may induce different subpopulations to change their treatment status and therefore estimate different LATEs.

[1] Latent outcomes are defined in Section 1.2. In general, latent outcomes will be functions of observed covariates. As is common in the literature, we keep this dependence implicit.

Hence, if a GMM overidentification test rejects, this no longer constitutes compelling evidence that one instrument is invalid. Rather, it might as well be interpreted as evidence for a non-constant treatment effect (Heckman, Schmierer, and Urzua 2010).

In this paper, we present an instrument test that is valid under essential heterogeneity. A key assumption of Imbens and Joshua Angrist 1994, which we maintain as well, is treatment monotonicity. Intuitively, this assumption says that individuals can be ordered by their willingness to participate in the treatment. As we show below, an immediate consequence of the monotonicity assumption is that the propensity score, i.e., the proportion of individuals who participate in the treatment, serves as an index that subsumes all information about observed outcomes that is included in a vector of instruments. We test the null hypothesis that this kind of index sufficiency holds, as this is a necessary and testable prerequisite for the intractable hypothesis of instrument validity. More concretely, we assume that a binary and a continuous instrument are available. The purpose of the binary instrument is to split the population into two subpopulations. We test whether observed outcomes conditional on the propensity score are identical in the two subpopulations. The reason why we assume continuity of the second instrument is that this offers a plausible way to argue that the supports of the propensity scores in the two subpopulations overlap.

Our test is related to the test of the validity of the matching approach suggested in Heckman et al. 1996 and Heckman et al. 1998. Their test also exploits index sufficiency under the null hypothesis. Moreover, the role that random assignment to a control group serves in their testing approach is similar to that of the binary instrument in our overidentification result. The testing theory that we develop in this paper translates with slight modifications to the testing problem of Heckman et al. 1996 and Heckman et al. 1998. We hope that it will prove useful in other settings where the null hypothesis imposes some kind of index sufficiency as well.

Our testable restriction in terms of a conditional mean function is closely related to a similar restriction in terms of the Marginal Treatment Effect (MTE, see Heckman and Vytlacil 2005 for a discussion of the MTE). The characterization of the restriction in terms of the MTE, while certainly the less practical one for testing, has a lot of theoretical appeal as it illustrates that our test is based on the overidentification of a structural parameter of the model.

We are not the first to consider the problem of testing instruments in a model with essential heterogeneity. Following previous work by Balke and Pearl 1997, Kitagawa 2013 and Huber and Mellace 2014 consider testing the validity of a discrete instrument in a LATE model. They test inequalities for the densities and the mean of the outcomes for always takers and never takers, i.e. two subpopulations for which treatment status is not affected by the instrument. In stark contrast, our test focuses on the subpopulation which responds to the instrument. Fernández-Val and Josh Angrist 2013 develop a LATE overidentification test under the additional assumption that the heterogeneity is captured by observed covariates. We do not require such an assumption. Our test lends itself naturally to testing continuous instruments, whereas previous tests can handle continuous instruments only via a discretization.

Our method works if both a binary and a continuous instrument are available. This is the case in many relevant applications. In this paper, we apply our method to test the validity of instruments that have been used to investigate the effect of teenage child bearing on high school completion. For another example of an evaluation problem where our method would come to bear, consider Carneiro, Heckman, and Vytlacil 2011. They estimate returns to schooling, using a binary indicator of distance to college, tuition fees, as well as a continuous measure of local labor market conditions as instruments.

Our test reduces to the problem of testing the equality of two nonparametric regression curves. This is a problem with a rich history in the statistical literature (cf., e.g., Hall and Hart 1990; King, Hart, and Wehrly 1991; Delgado 1993; Dette and Neumeyer 2001; Neumeyer and Dette 2003). Our testing problem, however, does not fit directly into any of the frameworks analyzed in the previous literature as it comes with the added complication of generated regressors. We propose a test statistic and quantify the effect of the first stage estimation error on the asymptotic distribution of the test statistic. We find that, in order to have good power against local alternatives we have to reduce the nonparametric bias from the first stage estimation. With our particular choice of second stage estimator, no further bias reduction is necessary.

We propose a bootstrap procedure to compute critical values. In the context of a treatment model with nonparametrically generated regressors, Y. Lee 2013 establishes the validity of a multiplier bootstrap that is based on the first order terms in an asymptotic expansion of the underlying process. We suggest a wild bootstrap procedure that does not rely on first order asymptotics and that is easy to implement in standard software. In exploratory simulations, our procedure is faithful to its nominal size in small and medium sized samples.

The paper is structured as follows: Section 1.2 defines our heterogeneous treatment model. In Section 1.3 we give an intuitive overview of our method, state our central overidentification result, discuss nonparametric parameter estimation, and define the test statistic. The asymptotic behavior of our test statistic is discussed in Section 1.4. Our simulations are presented in Section 1.5. In Section 1.6 we apply our approach to real data and study the validity of instruments in the context of teenage child bearing and high school graduation. Section 1.7 concludes.

1.2 Model definition

Our version of a treatment model with unobserved heterogeneity in the spirit of Imbens and Joshua Angrist 1994 is owed in large part to Vytlacil 2002. As in Abadie 2003 and Frölich 2007, we assume that our assumptions hold conditional on a set of covariates. We restrict ourselves to covariates that take values in a finite set. Our main overidentification result carries over to more general covariate spaces in a straightforward manner. The purpose of the restriction is exclusively to facilitate estimation by keeping the estimation of infinite dimensional nuisance parameters free from the curse of dimensionality. Without loss of generality, assume that we can enumerate all possible covariate configurations by $\{1, \ldots, J^{\max}\}$ and let J denote the covariate configuration of an individual. Treatment status is binary and is denoted by D. The latent outcomes are denoted by Y^0 and Y^1

and $Y = (1 - D)Y^0 + DY^1$ denotes the observed outcome. Note that by setting $\alpha = Y^0$ and $\beta = Y^1 - Y^0$, we recover the correlated random effects model from equation (1.1). Let S denote a continuous random variable and let Z denote a binary random variable. Below, S and Z are required to fulfill certain conditional independence assumptions that render them valid instruments in a heterogeneous treatment model. We observe a sample $(Y_i, D_i, S_i, Z_i, J_i)_{i \le n}$ from (Y, D, S, Z, J). Treatment status is determined by the threshold crossing decision rule

$$D = 1_{\{r_{Z,J}(S) \ge V\}},$$

with $r_{z,j}$ a function that is bounded between zero and one and V satisfying

$$V \sim U[0,1] \quad \text{and} \quad V \perp\!\!\!\perp (S, Z) \mid J. \tag{I-V}$$

Under this assumption, the function $r_{z,j}$ is a propensity score and V can be interpreted as an individual's type reflecting her natural proclivity to select into the treatment group. As pointed out in Vytlacil 2002, the threshold crossing model imposes treatment monotonicity.[2] The assumption that V is uniformly distributed is merely a convenient normalization that allows us to identify $r_{z,j}$. The crucial part of this assumption is that the instruments are jointly independent of the heterogeneity parameter V. This allows us to use the instruments as a source of variation in treatment participation that is independent of the unobserved types. Furthermore, we assume that for given V, Z and J the latent outcomes are independent of S,

$$Y^d \perp\!\!\!\perp S \mid V, Z, J \quad d = 0, 1. \tag{CI-S}$$

Also, for given V and J the latent outcomes are independent of Z,

$$Y^d \perp\!\!\!\perp Z \mid V, J \quad d = 0, 1. \tag{CI-Z}$$

Intuitively, these assumptions state that once the unobserved type is controlled for, the instruments are uninformative about latent outcomes. Note that we do not place any restrictions on the joint distribution of potential outcomes and V. Economically this means that unobserved characteristics, such as personal taste, that enter into the decision to participate in the treatment, are allowed to be correlated with the latent outcomes. The more commonly assumed instrument condition is

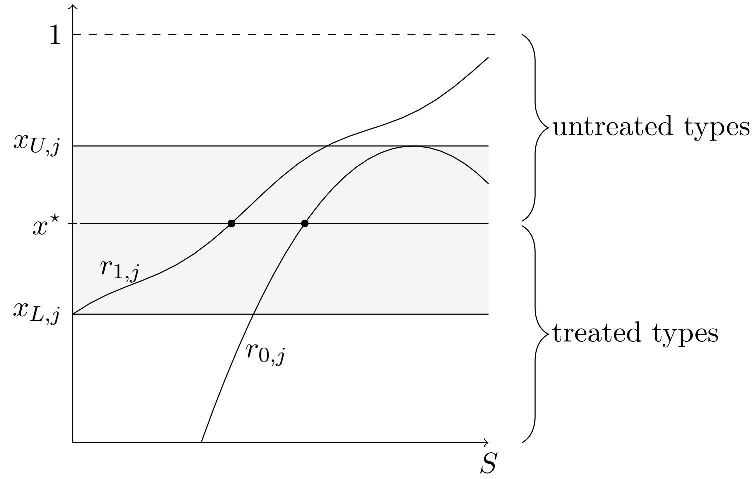
$$(Y^0, Y^1, V) \perp\!\!\!\perp (S, Z) \mid J,$$

which implies the conditional independence assumptions stated above. To argue the validity of an instrument it is helpful to split up the instrument condition in a way that allows us to disentangle participation and outcome effects. In our application, for example, assumptions CI-S and CI-Z seem quite plausible. The problematic assumption is to assume that the variation in treatment participation induced by the instrument is independent of the variation that is driven by the unobserved types.

Throughout, we let E_z and $\mathrm{E}_{z,j}$ denote the expectation operator conditional on $Z = z$, and $(J, Z) = (j, z)$, respectively.

[2] Consider two types $v_1 \le v_2$. Under the threshold model v_1 participates if v_2 participates. This is independent of the shape of the propensity score function. In particular, monotonicity of the propensity score function in its parameters is not required.

Figure 1.1: Heuristic description of method.



1.3 Overidentification test

1.3.1 Testing approach

Before we formally introduce the overidentification test, we give a heuristic description of our testing approach. Our test is based on comparing observed outcomes in the $Z = 0$ and $Z = 1$ subpopulations. For a fixed covariate configuration j, Figure 1.1 shows hypothetical plots for the propensity scores in the two subpopulations. The ranges of the two functions overlap so that there is an interval of participation probabilities that can be achieved in both subpopulations by manipulating the continuous instrument. The lower and upper bound of this interval are denoted by $x_{L,j}$ and $x_{U,j}$, respectively. Consider the participation probability $x^\star$ lying in this interval. Whenever the participation probability $x^\star$ is observed, all types $V \le x^\star$ will choose to participate in the treatment and all types $V > x^\star$ will abstain from seeking treatment. In other words, if we observe the same propensity score in two subpopulations, then all types will arrive at identical participation decisions regardless of which subpopulation they are selected into. The participation decision fixes which of the two latent outcomes we observe. Therefore, by fixing the propensity score and comparing observed outcomes between the two subpopulations, we are in fact comparing latent outcomes. Under the null hypothesis, latent outcomes behave identically in the two subpopulations since, by assumption, valid instruments do not affect latent outcomes. Consequently, for a given propensity score, observed outcomes should behave identically in the $Z = 0$ and $Z = 1$ subpopulations if the model is correctly specified. In particular,

$$\mathrm{E}[Y \mid Z = 0, r_{0,j}(S) = x^\star] = \mathrm{E}[Y \mid Z = 1, r_{1,j}(S) = x^\star].$$

In our approach we test this equality for different $x^\star$.

1.3.2 Overidentification result

For $z = 0,1$ and $j = 1,\dots,J^{\max}$ define $m_{z,j}(x) = \mathrm{E}_{z,j}[Y \mid r_{z,j}(S) = x]$. The propensity score is identified from

$$r_{z,j}(s) = \mathrm{E}_{z,j}[D \mid S = s]$$

and therefore $m_{z,j}$ is identified on the interior of the support of $r_{z,j}(S) \mid Z = z$. Our test is based on the following overidentification result.

Theorem 1.1 (Overidentification) *Fix $j \in \{1,\dots,J^{max}\}$ and suppose that conditional on $J = j$ x lies in the interior of the support of both $r_{0,j}(S) \mid Z = 0$ and $r_{1,j}(S) \mid Z = 1$. Then $m_{z,j}$ does not depend on z, i.e., $m_{0,j}(x) = m_{1,j}(x)$. Let $m_j(x)$ denote the common value for all j and x that satisfy the assumption.*

Proof

$$\begin{aligned} m_{z,j}(x) =& \mathrm{E}[Y \mid r_{z,j}(S) = x, Z = z, J = j] \\ =&(1-x)\,\mathrm{E}[Y^0 \mid r_{z,j}(S) = x, V > x, Z = z, J = j] \\ &+ x\,\mathrm{E}[Y^1 \mid r_{z,j}(S) = x, V \le x, Z = z, J = j] \\ =&(1-x)\,\mathrm{E}[Y^0 \mid V > x, Z = z, J = j] + x\,\mathrm{E}[Y^1 \mid V \le x, Z = z, J = j] \\ =&(1-x)\,\mathrm{E}[Y^0 \mid V > x, J = j] + x\,\mathrm{E}[Y^1 \mid V \le x, J = j] \end{aligned}$$

Now note that the right hand side does not depend on z. □

The result says that under the null hypothesis that the model is correctly specified, the parameter m_j can be identified from two different subpopulations. Under alternatives, the instruments have a direct effect on outcomes that is not mediated through the propensity score. The overidentification restriction has some power to detect such alternatives because in the two subpopulations, distinct values of the instrument vector are used to identify the same parameter.

Suppose that for $j = 1,\dots,J^{\max}$ there are $\underline{x}_j$ and $\bar{x}_j$, $\underline{x}_j \le \bar{x}_j$, and open sets $\mathcal{G}_j$ such that

$$\operatorname{supp} r_{0,j}(S) \mid Z = 0, J = j \cap \operatorname{supp} r_{1,j}(S) \mid Z = 1, J = j \supseteq \mathcal{G}_j \supseteq [\underline{x}_j \bar{x}_j].$$

Theorem 1.1 implies that on $[\underline{x}_j \bar{x}_j]$ we have

$$m_{0,j}(x) - m_{1,j}(x) = 0. \tag{1.2}$$

We are testing this equality. For the test to have some bite we need $[\underline{x}_j \bar{x}_j]$ to be non-empty. Intuitively, what is required is that for fixed Z the continuous instrument is strong enough to induce as many individuals to change their treatment status as would be swayed to change their participation decision by a change in Z while keeping S fixed. An important case where this is not possible is if Z is a deterministic function of S.

The basic idea of the overidentification result does not rely on the continuity of S. However, continuity of S is crucial as it offers a way to ensure that the common support of the propensity scores in the two subpopulations with $Z = 0$ and $Z = 1$ can plausibly have positive probability. For a given j we refer to an interval $[\underline{x}_j, \bar{x}_j]$ that satisfies the above condition as a testable subpopulation. It consists of a set of unobserved types that can be induced to select in and out of treatment by marginal changes in the continuous instrument regardless of the value of the binary instrument. Therefore the types in this interval are part of the complier population as defined in Joshua Angrist, Imbens, and Rubin 1996.

Theorem 1.1 is implied by the stronger result

$$\mathrm{E}_j[Y \mid S, Z] = \mathrm{E}_j[Y \mid r_{Z,j}(S)] \quad a.s.. \tag{1.3}$$

This says that conditional on covariates, the propensity score aggregates all information that the instruments provide about observed outcomes. In that sense, our approach can be interpreted as a test of index sufficiency that is similar in spirit to the test of the validity of the matching approach suggested in Heckman et al. 1996; Heckman et al. 1998. The equivalence (1.3) remains true if Y is replaced by a measurable function of Y. By considering different functions of Y, a whole host of testable restrictions can be generated. One implication, for example, is that a conditional distribution function is overidentified. In this paper, we only consider overidentified conditional mean outcomes and leave the obvious extensions to future research. Our testable restriction (1.2) is closely related to the marginal treatment effect (MTE)

$$\beta_j(x) = \mathrm{E}_j[Y^1 - Y^0 \mid V = x],$$

which has been proposed as a natural way to parameterize a heterogeneous treatment model (Heckman and Vytlacil 2005). In fact, $\beta_j(x) = \partial_x m_j(x)$. Since we are testing for overidentification of a function, we are also testing for overidentification of its derivative. If we were to base our test directly on the MTE instead of mean outcomes, we would not be able to detect alternatives where instruments are uncorrelated with the treatment effect β but have a direct effect on the base outcome α. Another advantage of our mean outcome approach over a test based on the MTE is that we avoid having to estimate a derivative. In our nonparametric setting, derivatives are much harder to estimate than conditional means. However, if the econometrician is not interested in a direct effect on the base outcome and if a large sample is available, it might be beneficial to look at β_j rather than at m_j. The reason is that as m_j is a smoothed version of β_j, it might not provide good evidence for perturbations of β_j that oscillate around zero. Another maybe more compelling reason to consider overidentification of β_j is that it allows us to investigate the source of a rejection of the null hypothesis. If a test based on m_j rejects while at the same time a test based on β_j does not reject, it seems likely that instruments have a direct effect on the base outcome but not on the treatment effect. In this paper, we focus on the test based on conditional outcomes and leave a test considering the MTE to future research.

It is helpful to think of alternatives as violations of the index sufficiency condition (1.3). Economically this means that instruments have a direct effect on outcomes, i.e., instruments have an effect on observed outcomes that can not be squared with their role as providers of independent variation in the participation stage. To formalize how our test detects such alternatives ignore covariates for the moment and define the prediction error from regressing on the propensity score instead of on the instruments

$$\varphi(S,Z) = \mathrm{E}[Y \mid S,Z] - \mathrm{E}[Y \mid r_Z(S)].$$

Now suppose that the model is correctly specified up to possibly a violation of the index sufficiency condition. The restricted null hypothesis is

$$H_0 : \varphi(S,Z) = 0 \quad a.s..$$

Using this notation we can rewrite the testable restriction (1.2) as

$$\mathrm{E}[\varphi(S,Z) \mid r_0(S) = x, Z = 0] - \mathrm{E}[\varphi(S,Z) \mid r_1(S) = x, Z = 1] = 0$$

for all $x \in [\underline{x}, \bar{x}]$. This is a necessary condition for

$$\mathrm{E}[\varphi(S,Z) \mid r_z(S) = x, Z = z] = 0 \quad \text{for } z = 0,1 \text{ and } x \in \operatorname{supp} r_z(S) \mid Z = z$$

which in turn is necessary for the restricted null. Since we are only testing a necessary condition not all alternatives can be detected. As an extreme case consider the case of identical propensity scores, i.e., $r_0 = r_1$. In this particular case our testable restriction does not have the power to detect a direct effect of S on outcomes.

1.3.3 Parameter estimation and test statistic

Let $\hat{m}_{z,j}$ denote an estimator of $m_{z,j}$ and let $\underline{x} = (\underline{x}_1, \dots, \underline{x}_{J^{\max}})$ and $\bar{x} = (\bar{x}_1, \dots, \bar{x}_{J^{\max}})$. Suppose that under the null hypothesis m_j is overidentified on $[\underline{x}_j, \bar{x}_j]$ for $j = 1, \dots, J^{\max}$ and define the test statistic

$$T_n = T_n(\underline{x}, \bar{x}) = \sum_{j=1}^{J} \int_{\underline{x}_j}^{\bar{x}_j} (\hat{m}_{0,j}(x) - \hat{m}_{1,j}(x))^2 \pi_j(x)\, dx. \tag{1.4}$$

Here π_j is a weight function that can be used to fine-tune power against certain alternatives. What constitutes a sensible choice for π_j will depend on the specifics of the application. For simplicity we assume that π_j is unity from here on. In the following we will refer to the subsample with $J_i = j$ and $Z_i = z$ as the (j,z)-cell. We estimate $\hat{m}_{z,j}$ by a two step procedure. In the first step we estimate the function $r_{z,j}$ by local polynomial regression of S on D within the (j,z)-cell. We will refer to this step as the participation regression. The first step estimator is denoted by $\hat{r}_{z,j}$. In the second step we estimate $m_{z,j}$ by local linear regression of Y on the predicted regressors $\hat{r}_{z,j}(S_i)$ within the (j,z)-cell. This step will be referred to as outcome regression. We let L and K denote the kernel functions for the participation and outcome regression, respectively. Also let g and h denote the respective bandwidth sequences. To reduce notational clutter, we assume that the bandwidths do not depend on j and z. It is straightforward to extend the model to allow cell dependent bandwidths. Let q denote the degree of the local polynomial in the participation regression. It is necessary to choose $q \geq 2$ to remove troublesome bias terms.

If these bias terms are not removed the test will behave asymptotically like a linear test, i.e., it will favor the rejection of alternatives that point into a certain direction. A formal definition of the estimators is provided in Appendix 1.A.

In many applications the bounds $\underline{x}$ and $\bar{x}$ are not a priori known and have to be estimated. Below we show that replacing the bounds by a consistent estimator does not affect the asymptotic distribution of the test statistic under weak assumptions. Since we assume $r_{z,j}$ to be continuous, the set on which m_j is overidentified will always be an interval $(x_{L,j}, x_{U,j})$. To avoid boundary problems we fix some positive c_δ and estimate the smaller interval $[\underline{x}_j, \bar{x}_j] = [x_{L,j} + c_\delta, x_{U,j} - c_\delta]$ by its sample equivalent.

1.3.4 Inference and bootstrap method

In Theorem 1.2 below we characterize the asymptotic distribution of the test statistic under the null. However, as we explain below, we do not recommend to use this distributional result as a basis for approximating critical values. In a related problem with nonparametrically generated regressors Y. Lee 2013 establishes the validity of a multiplier bootstrap procedure. We conjecture that, building on the asymptotic influence function from Lemma 1.3 in the appendix, a similar approach can be taken in our setting. However, simulating the distribution by multiplier methods has some disadvantages. First, as the approach is based on asymptotic influence functions no improvements beyond first order asymptotics can be expected. Secondly, the method requires significant coding effort which makes it unattractive in applied work. This is why we propose a wild bootstrap procedure that is straightforward to implement instead. We provide simulation evidence that illustrates that the procedure can have good properties in small and medium sized samples. A theoretical proof of the validity of the method is beyond the scope of the present paper and left to future research.

First, estimate the bounds $\underline{x}$ and $\bar{x}$. In the bootstrap samples these bounds can be taken as given. For all j and all z estimate $\hat{r}_{z,j}$ from the (j,z)-cell and predict $R_i^0 = \hat{r}_{Z_i,J_i}(S_i)$ and $\hat{\zeta}_i^0 = D_i - R_i^0$. Next, pool all observations with $J = j$ and estimate m_j by local linear regression of Y_i on R_i^0 with kernel K and bandwidth h. Predict $M_i^0 = \hat{m}_{J_i}(R_i^0)$ and $\hat{\epsilon}_i^0 = Y_i - M_i^0$. Now generate B bootstrap samples in the following way. Draw a sample of n independent Rademacher random variables $(W_i)_{i\leq n}$, let

$$\begin{pmatrix} D_i^* \\ Y_i^* \end{pmatrix} = \begin{pmatrix} R_i^0 \\ M_i^0 \end{pmatrix} + W_i \begin{pmatrix} \hat{\zeta}_i^0 \\ \hat{\epsilon}_i^0 \end{pmatrix},$$

and define the bootstrap sample $(Y_i^*, D_i^*, S_i, Z_i, J_i)_{i\leq n}$.

While we use Rademacher variables as an auxiliary distribution, other choices such as the two-point distribution from Mammen 1993 or a standard normal distribution are also possible.

1.4 Asymptotic analysis

In this section we derive the asymptotic distribution of our test statistic. This analysis gives rise to a number of interesting insights. First, it allows us to consider local alternatives. A lesson implicit in the existing literature on L^2-type test statistics is that a naive construction of such a statistic often leads to a test with the undesirable property of treating different local alternatives disparately. Loosely speaking, such a tests behaves like a linear test in that it only looks for alternatives that point to the same direction as a certain bias term (cf. Härdle and Mammen 1993). We find that in order to avoid such behavior it suffices to employ bias-reducing methods when estimating the propensity scores. We recommend to fit a local polynomial of at least quadratic degree. The outcome estimation does not contribute to the problematic bias term. Secondly, our analysis allows us to consider the case when the bounds of integration $\underline{x}$ and $\bar{x}$ are unknown and have to be estimated. We show that, provided that the estimators satisfy a very weak assumption, the asymptotic distribution is unaffected by the estimation. Thirdly, our results allow us to make recommendations about the choice of the smoothing parameters. Our main asymptotic result implies that our test has good power against a large class of local alternatives if the outcome stage estimator oversmoothes compared to the participation stage estimator but not by too much. For convenience of notation, in the following we focus on the case $J^{\max} = 1$ and omit the j subscript. Proofs for the results in this section can be found in the appendix.

1.4.1 Assumptions

Define the sampling errors $\epsilon = Y - \mathrm{E}[Y \mid r_Z(S)]$ and $\zeta = D - E[D \mid S, Z]$. Under the null hypothesis the conditional variances $\sigma^2_\epsilon(x) = \mathrm{E}[\epsilon^2 \mid r_Z(S) = x]$, $\sigma^2_\zeta(x) = \mathrm{E}[\zeta^2 \mid r_Z(S) = x]$ and $\sigma_{\epsilon\zeta}(x) = \mathrm{E}[\epsilon\zeta \mid r_Z(S) = x]$ remain unchanged if the unconditional expectation operator is replaced by the conditional expectation operator E_z, $z = 0, 1$. Also note that $\sigma^2_\zeta(x) = x(1-x)$. For our local estimation approach to work we have to impose some smoothness on the functions m_z and r_z. We now give conditions in terms of the primitives of the model to ensure that the functions that we are estimating are sufficiently smooth.

Assumption 1.1
Assume that m is overidentified on an open interval (x_L, x_U) and

(i) there is a positive ρ such that

$$\mathrm{E}[\exp(\rho |Y^d|)] < \infty, \quad d = 0, 1.$$

(ii) Conditional on $Z = z$, $z = 0, 1$, S is continuously distributed with density $f_{S|Z=z}$ and $r_z(S)$ is continuously distributed with density $f_{R|Z=z}$. Moreover, $f_{S|Z=z}$ is bounded away from zero and has one bounded derivatives and $f_{R|Z=z}$ is bounded away from zero and is twice continuously differentiable.

(iii) $\mathrm{E}[Y^0 \mid V > x]$ and $\mathrm{E}[Y^1 \mid V \le x]$ are twice continuously differentiable on (x_L, x_U).

(iv) The functions $\mathrm{E}[(Y^0)^2 \mid V > x]$ and $\mathrm{E}[(Y^1)^2 \mid V \le x]$ are continuous on (x_L, x_U).

(v) r_z, $z = 0, 1$, is $(q+1)$-times continuously differentiable on (x_L, x_U).

The assumption implies standard regularity conditions for m, σ_ϵ^2 and $\sigma_{\epsilon\zeta}$ that are summarized in Assumption 1.3 in the appendix. These conditions include that m is twice continuously differentiable and that σ_ϵ^2 and $\sigma_{\epsilon\zeta}$ are continuous. A consequence of Assumption 1.1(ii) is that x_L and x_U are identified by

$$\begin{aligned} x_L &= \max\left\{\inf_s r_0(s), \inf_s r_1(s)\right\} \quad \text{and} \\ x_U &= \min\left\{\sup_s r_0(s), \sup_s r_1(s)\right\}. \end{aligned} \tag{1.5}$$

Fix a small constant $c_\delta > 0$. We can choose $\underline{x} = x_L + c_\delta$ and $\bar{x} = x_U - c_\delta$. We also need some assumptions about the kernel functions.

Assumption 1.2
K and L are symmetric probability density functions with bounded support. K has two bounded and continuous derivatives. The bandwidth sequences are parametrized by $g \sim n^{-\eta^*}$ and $h \sim n^{-\eta}$.

Implicit in this assumption is that the bandwidths are not allowed to depend on z. In particular, the bandwiths are tied to the overall sample size rather than the size of the two subsamples corresponding to $Z = z$, $z = 0, 1$. This is for expositional convenience only.

1.4.2 Local alternatives

To investigate the behavior of the test under local alternatives we now consider a sequence of models that converges to a model in the null hypothesis.

Definition 1.1 (Local alternative) *A sequence of local alternatives is a sequence of models*

$$\mathcal{M}^n = (Y^{0,n}, Y^{1,n}, V^n, S, Z, r_0, r_1)$$

in the alternative that converges to a model

$$\mathcal{M}^{null} = (Y^{0,null}, Y^{1,null}, V^{null}, S, Z, r_0, r_1)$$

in the null hypothesis in the following sense:

$$\sup_x \mathrm{E}\left[\left(1_{\{V^n \le x\}} - 1_{\{V^{null} \le x\}}\right)^2 \mid S, Z\right] = O_{a.s}\left(c_n^2\right) \tag{1.6a}$$

$$\mathrm{E}\left[\left(Y^{d,n} - Y^{d,null}\right)^2 \mid S, Z\right] = O_{a.s}\left(c_n^2\right) \quad d = 0, 1 \tag{1.6b}$$

for a vanishing sequence c_n. For n large enough there are positive constants ρ and C such that

$$\mathrm{E}[\exp(\rho\left|Y^{d,n} - \mathrm{E}[Y^{d,n} \mid S, Z]\right|) \mid S, Z] \le C \quad d = 0, 1.$$

We let Y^n and Y^{null} denote the observed outcome under the model $\mathcal{M}^n$ and $\mathcal{M}^{null}$, respectively.

Write φ_n for the index prediction error under the sequence of models $\mathcal{M}^n$ and note that

$$\begin{aligned}\varphi_n(S,Z) &= \mathrm{E}[Y^n \mid S,Z] - \mathrm{E}[Y^n \mid r_Z(S)] \\ &= \mathrm{E}[Y^n - Y^{\text{null}} \mid S,Z] - \mathrm{E}[Y^n - Y^{\text{null}} \mid r_Z(S)] = O_{a.s}(c_n)\end{aligned}$$

so that index sufficiency holds approximately in large samples. Formally, we are testing the sequence of local alternatives

$$H_{0,n} : \Delta_n(x) = 0 \quad \text{for } x \in [\underline{x}, \bar{x}]$$

with

$$\Delta_n(x) = \mathrm{E}[\varphi_n(S,Z) \mid r_Z(S) = x, Z = 0] - \mathrm{E}[\varphi_n(S,Z) \mid r_Z(S) = x, Z = 1].$$

To analyze the behavior of our test under local alternatives we suppose that we are observing a sequence of samples where the n-th sample is drawn from $\mathcal{M}^n$. For vanishing c_n we interpret $\mathcal{M}^{\text{null}}$ as a hypothetical data generating process that satisfies the restriction of the null and that is very close to the observed model $\mathcal{M}^n$. Our objective is to show that our test can distinguish $\mathcal{M}^n$ from $\mathcal{M}^{\text{null}}$. The fastest rate at which local alternatives can be detected is $c_n = n^{-1/2}h^{-1/4}$. This is the standard rate for this type of problem (cf. Härdle and Mammen 1993). At this rate the smoothed and scaled version of the local alternative

$$\Delta_{K,h}(x) = c_n^{-1} \int \Delta_n(x + ht)K(t)\, dt$$

enters the asymptotic distribution of the test statistic.

1.4.3 Asymptotic behavior of the test statistic

For our main asymptotic result below we use the asymptotic framework introduced in the previous subsection where T_n is the test statistic computed on a sample of size n drawn from the model $\mathcal{M}^n$. The result states that the asymptotic distribution of the test statistic can be described by the asymptotic distribution of the statistic under the hypothetical model $\mathcal{M}^{\text{null}}$ shifted by a deterministic sequence that measures the distance of the observed model $\mathcal{M}^n$ from $\mathcal{M}^{\text{null}}$. The behavior of the test statistic under the null is obtained as a special case by choosing a trivial sequence of local alternatives.

Theorem 1.2 *Let $c_n = n^{-1/2}h^{-1/4}$ and consider a model $\mathcal{M}^{null}$ satisfying Assumption 1.1 for $x_L < \underline{x} < \bar{x} < x_U$ and corresponding local alternatives $\mathcal{M}^n$ satisfying Definition 1.1. The functions $\mathrm{E}[Y^n \mid r_Z(S) = x]$ and $\mathrm{E}[Y^n \mid r_Z(S) = x, Z = z]$, $z = 0,1$, are Riemann integrable on (x_L, x_U). The bandwidth parameters η and η^* satisfy*

$$3\eta + 2\eta^* < 1 \qquad (1.7a) \qquad\qquad \eta > 1/6 \qquad (1.7d)$$

$$2\eta > \eta^* \qquad (1.7b) \qquad\qquad (q+1)\eta^* > 1/2 \qquad (1.7e)$$

$$\eta^* + \eta < 1/2 \qquad (1.7c) \qquad\qquad \eta^* > \eta\,. \qquad (1.7f)$$

Then

$$n\sqrt{h}T_n - \frac{1}{\sqrt{h}}\gamma_n - \int_{\underline{x}}^{\bar{x}} \Delta_{K,h}^2(x)\,dx \xrightarrow{d} N(0,V),$$

where

$$V = 2K^{(4)}(0)\int_{\underline{x}}^{\bar{x}} \left[x(1-x)m'(x)^2 - 2\sigma_{\epsilon\zeta}(x)m'(x) + \sigma_\epsilon^2(x)\right]^2 \left(\sum_{z\in\{0,1\}} \frac{1}{p_z f_{R,z}(x)}\right)^2 dx$$

and γ_n is a deterministic sequence such that $\gamma_n \to \gamma$ for

$$\gamma = K^{(2)}(0)\int_{\underline{x}}^{\bar{x}} \left[x(1-x)m'(x)^2 - 2\sigma_{\epsilon\zeta}(x)m'(x) + \sigma_\epsilon^2(x)\right] \sum_{z\in\{0,1\}} \frac{1}{p_z f_{R,z}(x)}\,dx.$$

Here $m(x) = E[Y^{null} \mid r_Z(S) = x]$ and the conditional covariances are computed under $\mathcal{M}^{null}$. $K^{(v)}$ denotes the v-fold convolution product of K. For $q \geq 2$ the set of admissible bandwidths is non-empty.

The result implies that the test can detect local alternatives that converge to a model in the null hypothesis at the rate $c_n = n^{-1/2}h^{-1/4}$ and that satisfy

$$\liminf_n \int_{\underline{x}}^{\bar{x}} \Delta_{K,h}^2(x)\,dx > 0.$$

Both the first and the second stage estimation contribute to the asymptotic variance. The term $x(1-x)m'(x)^2 - 2\sigma_{\epsilon\zeta}(x)m'(x)$ in the expression for the asymptotic variance is due to the first stage estimation. Under our assumptions this term can not be signed, so that the first stage estimation might increase or decrease the asymptotic variance. However, while it is possible to construct models under which this term is negative, these models have some rather unintuitive features and we do not consider them to be typical. If the estimated regression function is rather flat, the influence of the first stage regression on the asymptotic variance is small. To gain an intuition as to why this is so, note that if $m'(x)$ is small then a large interval of index values around x is informative about $m(x)$. This helps to reduce the first stage estimation error, because on average the index is estimated more reliably over large intervals than over smaller intervals.

An essential ingredient in the proof of Theorem 1.2 is a result from Mammen, Rothe, and Schienle 2012. They provide a stochastic expansion of a local linear smoother that regresses on generated regressors around the oracle estimator. The oracle estimator is the infeasible estimator that regresses on the true instead of the estimated regressors. This expansion allows us to additively separate the respective contributions of the participation and the outcome regression to the overall bias of our estimator of $m_0 - m_1$. Under the null the oracle estimator is free of bias. This is intuitive. Under the null, $m = m_0 = m_1$ so that $\hat{m}_0$ and $\hat{m}_1$ estimate the same function in two subpopulations with non-identical designs. A well-known property of the local linear estimator is that its bias is design independent (Ruppert and Wand 1994) which makes it attractive for testing problems that compare nonparametric fits (Gørgens 2002). Hence, only the bias of the participation regression has to be reduced.

We do not recommend using the distributional result in Theorem 1.2 to compute critical values. The exact shape of the distribution is very sensitive to bandwidth choice. As explained below, one does not know in practice if bandwidths satisfy the conditions in the theorem. Even if bandwidths are chosen incorrectly, in many cases the statistic still converges to a normal and most of the lessons we draw from the asymptotic analysis still hold up. However, the expressions for the asymptotic bias and variance would look different. Furthermore, to estimate the asymptotic bias and variance we have to estimate derivatives and conditional variances. These are quantities that are notoriously difficult to estimate. Instead, our inference is based on the wild bootstrap procedure introduced in Section 1.3. We investigate the validity of our bootstrap procedure in simulations in Section 1.5 below.

Theorem 1.2 requires that the bandwidth parameters satisfy a system of inequalities. The restrictions are satisfied for example if $q = 2$, $\eta^* = 1/5$ and $1/6 < \eta < 1/5$. The inequalities (1.7a)-(1.7c) ensure that our estimators satisfy the assumptions of Theorem 1 in Mammen, Rothe, and Schienle 2012. Condition (1.7d) ensures that up to parametric order the bias of the oracle estimator is design independent. When the inequality (1.7f) is satisfied, the error terms from both the participation and outcome regression contribute to the asymptotic distribution. Finally, inequality (1.7e) says that the bias from the participation regression must vanish at a faster than parametric rate. This is precisely the condition needed to get rid of the troublesome bias terms discussed above. While the proposition offers conditions on the rates at which the bandwidths should vanish it offers little guidance on how to choose the bandwidths in finite samples. There are no bandwidth selection procedures that produce deliberately under- or oversmoothing bandwidths. This problem is by no means specific to our model but on the contrary quite ubiquitous in the kernel smoothing literature (cf. Hall and Horowitz 2012). In our application we circumvent the problem of bandwidth selection by reporting results for a large range of bandwidth choices.

In practice, the bounds of integration $\underline{x}$ and $\bar{x}$ are additional parameters that have to be chosen. In most applications, this means that they have to be estimated from the data. The following result states that a rather slow rate of convergence of these estimated bounds suffices to ensure that bound estimation does not affect the asymptotic distribution.

Theorem 1.3 *Suppose that the assumptions of Theorem 1.2 hold. Assume also that* $\underline{x}_n$ *and* $\bar{x}_n$ *are sequences of random variables such that*

$$(\underline{x}_n, \bar{x}_n) - (\underline{x}, \bar{x}) = o_p\left(h^{\ell}\right)$$

for a constant $\ell > 1/2$. *Then*

$$T_n(\underline{x}_n, \bar{x}_n) - T_n(\underline{x}, \bar{x}) = o_p\left(\frac{1}{n\sqrt{h}}\right).$$

Let $\hat{x}_L$ and $\hat{x}_U$ denote the sample equivalents of the right hand side of the equation (1.5) that identifies x_L and x_U, respectively. Under the bandwidth restrictions of Theorem 1.2 the assumptions in Theorem 1.3 are satisfied if we set $\underline{x}_n = \hat{x}_L + c_\delta$ and $\bar{x}_n = \hat{x}_U - c_\delta$.

alternative	perturbation
1	$\Delta_\alpha = 0.2$
2	$\Delta_\alpha = -\frac{1}{2}V$
3	$\Delta_\alpha = 40(V - 0.3)\exp\left(-80(V - 0.3)^2\right)$
4	$\Delta_\beta = 0.2$
5	$\Delta_\beta = -V$
6	$\Delta_\beta = 40(V - 0.3)\exp\left(-80(V - 0.3)^2\right)$

Table 1.1: Specification of simulated alternatives.

1.5 Simulations

We simulate various versions of the random coefficient model from equation (1.1) and compute empirical rejection probabilities for our bootstrap test for two sample sizes and a large number of bandwidth choices. As in the previous section we assume $J^{\max} = 1$ and drop the j subscript.

Our basic setup is a model in the null hypothesis. Simulating our test for this model allows us to compare the nominal and empirical size of our test. We then generate several models in the alternative by perturbing outcomes in the basic model for the $Z = 1$ subpopulation. For the basic model we define linear propensity scores $r_0(s) = 0.1{+}0.5s$ and $r_1(s) = 0.5s$. The binary instrument Z is a Bernoulli random variable with $P(Z = 0) = P(Z = 1) = 0.5$ and the continuous instrument S is distributed uniformly on the unit interval. The base outcome α follows a mean-zero normal distribution with variance 0.5. The treatment effect is a deterministic function of V, $\beta = -2V$. As alternatives we consider perturbations of the base outcome α as well as perturbations of the treatment effect β. These perturbations are obtained by adding Δ_α to α and Δ_β to β in the $Z = 1$ subpopulation. The specifications for the alternatives are summarized in Table 1.1. The first three alternatives consider perturbations of the base outcome, whereas alternatives 4-6 are derived from perturbations of the treatment effect. Alternatives 1 and 4 consider the case that base outcome and treatment effect, respectively, are shifted independently of the unobserved heterogeneity V. The perturbations generating alternatives 2 and 5 are linear functions of V. Finally, alternatives 3 and 6 are generated by perturbing by functions of V that change sign. These alternatives are expected to be particularly hard to detect because our test is based on the m_z function which smoothes over the unobserved heterogeneity as is apparent in the proof of Proposition 1.1. As bandwidths we choose $g = C_g n^{-\frac{1}{5}}$ and $h = C_h n^{-\frac{1}{6}}$. We report results for a number of choices for the constants C_g and C_h. We set $q = 2$ and choose an Epanechnikov kernel for both K and L. The sample size is set to $n = 200, 400$. These should be considered rather small numbers considering the complexity of the problem. We consider the nominal levels $\theta = 0.1, 0.05$ as these are the most commonly used ones in econometric applications. As bound estimation has only a higher-order effect we take $\underline{x} = 0.15$ and $\bar{x} = 0.45$ as given. To simulate the bootstrap distribution we are using $B = 999$ bootstrap iterations. For

	$\theta = 0.10$						$\theta = 0.05$					
C_h	0.50	0.75	1.00	1.25	1.50	1.75	0.50	0.75	1.00	1.25	1.50	1.75
null												
$C_g = 0.50$	9.3	8.9	8.4	7.7	8.6	9.6	4.2	3.4	4.7	4.2	4.1	4.1
$C_g = 0.75$	10.1	9.9	9.4	8.2	7.7	9.3	4.8	4.5	4.0	3.3	3.6	4.0
$C_g = 1.00$	8.9	8.7	7.4	9.0	8.9	8.1	4.2	4.1	3.2	4.0	3.6	3.3
alternative 1												
$C_g = 0.50$	94.3	93.8	93.7	93.6	92.8	94.7	88.5	87.1	87.2	86.7	87.7	88.4
$C_g = 0.75$	94.8	91.9	93.0	92.6	94.0	93.8	88.6	86.9	87.2	85.8	87.3	87.2
$C_g = 1.00$	94.0	93.4	94.8	93.5	93.8	93.3	86.7	88.4	89.6	87.2	87.2	89.3
alternative 2												
$C_g = 0.50$	96.9	97.5	97.5	98.1	98.6	98.0	93.3	94.4	95.4	96.0	96.4	95.4
$C_g = 0.75$	96.9	97.9	97.2	97.8	97.1	97.5	93.0	95.6	94.6	94.7	94.3	95.3
$C_g = 1.00$	97.7	97.2	97.4	97.8	97.4	97.8	94.5	95.3	94.1	94.1	95.3	94.4
alternative 3												
$C_g = 0.50$	8.3	8.7	7.2	9.3	8.7	8.9	3.4	3.6	3.5	4.6	4.0	4.0
$C_g = 0.75$	6.9	9.1	8.9	8.6	8.9	9.3	3.5	4.4	3.6	3.6	4.0	3.9
$C_g = 1.00$	8.3	8.2	7.9	8.8	8.9	8.7	4.0	3.7	3.7	3.7	4.6	3.9
alternative 4												
$C_g = 0.50$	25.5	23.8	22.9	24.2	22.6	22.7	15.1	13.9	13.5	13.8	12.3	13.3
$C_g = 0.75$	25.1	26.3	26.1	22.3	23.3	24.7	15.0	14.6	15.0	13.1	13.3	14.5
$C_g = 1.00$	25.4	23.5	24.5	23.7	23.7	23.6	15.2	13.0	15.6	13.8	14.1	13.8
alternative 5												
$C_g = 0.50$	24.3	22.5	21.5	22.7	22.8	21.8	14.9	12.9	11.9	13.8	12.0	12.4
$C_g = 0.75$	21.1	22.0	21.3	20.9	22.7	22.3	10.4	10.8	12.1	11.5	12.7	12.5
$C_g = 1.00$	21.8	21.5	21.4	23.7	21.9	22.5	13.1	12.0	11.3	12.7	12.6	12.2
alternative 6												
$C_g = 0.50$	45.1	44.3	42.3	45.2	46.6	47.7	30.9	30.7	29.3	31.2	35.0	31.2
$C_g = 0.75$	45.3	43.5	44.9	44.2	45.8	44.2	32.3	31.4	32.0	30.7	33.0	30.3
$C_g = 1.00$	44.2	45.5	47.6	44.3	44.6	46.6	32.6	33.7	34.0	30.6	30.9	33.9

Table 1.2: Empirical rejection probabilities in percentage points under nominal level θ. Sample size is $n = 400$.

each model we conduct 999 simulations. Empirical rejection probabilities are reported in Table 1.2 for $n = 400$ and in Table 1.3 in the appendix for $n = 200$.

We discuss only the results for $n = 400$ in detail. Under the null hypothesis the empirical rejection probabilities are very close to the nominal levels. While this is not conclusive evidence that our bootstrap approach will always work, it is suggestive of the validity of the procedure.

Alternative 1 and Alternative 2 are detected with high probability. These alternatives are particularly easy to detect for two reasons. First, the perturbation affects a large subpopulation so that the alternative is easy to detect due to abundance of relevant data. Secondly, the smoothing inherent in the quantities that our test considers does not smear out the perturbations in a way that makes the alternatives hard to detect. To understand the first effect contrast Alternative 1 and Alternative 2 with Alternative 4 and

Alternative 5. Both pairs of alternatives arise from similar perturbations. However, the whole subsample with $Z = 1$ can be used to detect the first pair. In contrast, only treated individuals in the $Z = 1$ subsample provide data that helps to detect the second pair. A back-of-the-envelope calculation reveals that on average only about $400 \times 1/2 \times 1/4 = 25$ observations fall into the subsample with $Z = 1$ and $D = 1$. As cell sizes are observed in applications, a lack of relevant data is a problem that can readily be accounted for when interpreting test results. To shed light on the second effect recall that m_z is derived from smoothing outcomes over $V \leq x$ and $V > x$. Therefore, if a perturbation changes sign, positive and negative deviations from the null will cancel each other out. This effect is precisely what makes it so hard to detect perturbations such as those underlying Alternative 3 and Alternative 6. Luckily, these kinds of alternatives are not what should be expected in many applications. The problem that applied researchers have in mind most of the time is that instruments might have a direct effect on outcomes that can readily be signed by considering the economic context. In that respect, Alternative 1 and Alternative 2 are more typical of issues that applied economists worry about than Alternative 3.

It might seem puzzling that Alternative 6 is detected much more frequently than Alternative 3. The reason is that in Alternative 3 negative deviations in the $V \leq x$ population are offset by positive deviations in the $V > x$ population. This does not happen in Alternative 6 as only the treated population is affected by the perturbation.

Accounting for the complexity of the problem the sample size $n = 200$, for which we report results in the appendix, should be considered very small. Therefore, it is not surprising that the deviations from the nominal size are slightly more pronounced than in the larger sample. The deviations err on the conservative side, but that might be a particularity of our setup. The pattern in the way alternatives are detected is similar to the $n = 400$ sample with an overall lower detection rate.

Our simulations show that our approach has good empirical properties in finite samples. For the simulated model the test holds its size which indicates that the bootstrap procedure works well. Very particular alternatives that perturb outcomes by a function of the unobserved types that oscillates around zero are difficult to detect by our procedure. Alternatives that we consider to be rather typical are reliably detected provided that the subsample affected by the alternative is large enough.

1.6 Application

To illustrate the applicability of our method we now consider the effect of teenage child-bearing on the mother's probability of graduating from high-school. This topic has been discussed extensively in the literature. An early survey can be found in Hoffman 1998. To deal with the obvious endogeneity of motherhood, many authors (Ribar 1994; Hotz, McElroy, and Sanders 2005; Klepinger, Lundberg, and Plotnick 1995) have used instrumental variables methods. It has been suggested that treatment effect heterogeneity is a reason why estimated effects depend strongly on the choice of instrument (Reinhold 2007). In fact, it is very natural to assume that the effect of motherhood on graduation is heterogeneous. For a simple economic model that generates treatment effect heterogeneity suppose that the time cost of child care is the same for students of different abilities whereas the time cost of studying to improve the odds of graduating is decreasing in ability. To translate the problem into our heterogeneous treatment model let D denote a binary indicator of teenage motherhood and let Y denote a binary indicator of whether the woman has obtained a high school diploma[3]. We consider two instruments from the literature. The first one, henceforth labelled S, is age at first menstrual period which has been used in the studies by Ribar 1994 and Klepinger, Lundberg, and Plotnick 1995. This instrument acts as a random shifter of female fecundity and is continuous in nature. Its validity is discussed briefly in Klepinger, Lundberg, and Plotnick 1995 and Levine and Painter 2003. The second instrument, denoted by Z, is an indicator of whether the individual experienced a miscarriage as a teenager. Miscarriage has been used as an unexpected fertility shock in the analysis of adult fertility choices (Miller 2011) and also to study teenage child bearing in Hotz, Mullin, and Sanders 1997; Hotz, McElroy, and Sanders 2005. The population studied in Hotz, McElroy, and Sanders 2005 consists of all women who become pregnant in their teens, whereas we focus on the larger group of all women who are sexually active in their teens. This turns out to be a crucial difference. It stands to investigate the plausibility of the assumptions I-V, CI-S and CI-Z. Arguably, age at first menstrual period is drawn independently of V and fulfills the instrument specific conditional independence assumption CI-S if one controls for race. Possible threats to a linear version of CI-Z are discussed in Hotz, Mullin, and Sanders 1997. Hotz, McElroy, and Sanders 2005 conclude that the linear version of CI-Z holds in good approximation in the population that they are considering. The most problematic assumption to maintain is that Z is orthogonal to V. In a simplified behavioral model teenagers choose to become pregnant based on their unobserved type and then a random draw from nature determines how that pregnancy is resolved. This implies a sort of maximal dependence between Z and V, i.e., teenagers select into treatment and into $Z = 1$ in exactly the same way. Our test substantiates this heuristic argument by rejecting the null hypothesis that the assumptions I-V, CI-S and CI-Z hold simultaneously. Furthermore, it gives instructive insights into the role that heterogeneity plays in the failure of the assumptions.

We use data from the National Longitudinal Survey of Youth 1997[4] (henceforth NLSY97) from round 1 through round 15. We only include respondents who were at least 21 of age at the last interview they participated in. This is to ensure that we capture our outcome variable. A miscarriage is defined as a teenage miscarriage if the woman experiencing the miscarriage was not older than 18 at the time the pregnancy ended. Similarly, a young woman is defined as a teenage mother if she was not older than 18

[3] We do not include equivalency degrees (GED's). There is a discussion in the literature as to what the appropriate measure is (cf. Hotz, McElroy, and Sanders 2005).

[4] Most of the previous studies relied on data from the National Longitudinal Survey of Youth 1979 (NLSY79). In that study the date of the first menstrual period was asked for for the first time in 1984 when the oldest respondents were 27 years old. As is to be expected, a lot of respondents had trouble recalling the date such a long time after the fact. The NLSY97 contained the relevant question starting from the very first survey when the oldest respondents were still in their teens. Since our method relies on a good measurement of the continuous variable the NLSY97 data is a better choice than the NLSY79 data.

when the child was born. We control for race for two reasons. First, this is required to make the menarche instrument plausible. Secondly, this takes care of the oversampling of minorities in the NLSY97 so that we are justified in using unweighed estimates. We remove respondents who report “mixed race” as race/ethnicity because the cell size is too small to conduct inference. Table 1.4 in the appendix gives some summary statistics for our sample. An unfortunate side effect of using the low probability event of a teenage miscarriage as an instrument is that cell sizes can become rather small. This makes it impossible to control for additional covariates while preserving reasonable power. In Section 1.7 we briefly discuss a model that permits a much larger number of covariates. The estimated propensity scores $\hat{r}_{z,j}$ are plotted in Figure 1.2. For each j the functions

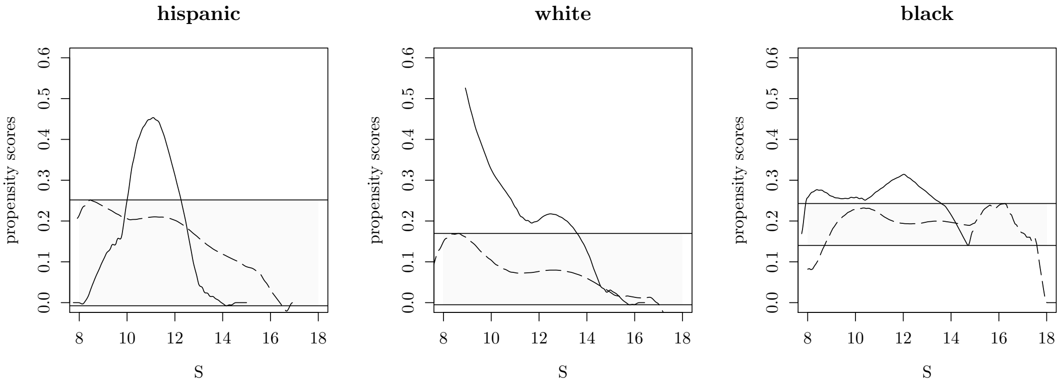


Figure 1.2: Probability of entering treatment conditional on age of first menstrual period (S) plotted separately for the subpopulations with $Z = 0$ (no miscarriage as a teenager, *dashed line*) and $Z = 1$ (miscarriage as a teenager, *solid line*). Plotted with $q = 1$ and bandwidth $g = 2.00$.

$\hat{r}_{0,j}$ and $\hat{r}_{1,j}$ are not identical almost everywhere and their ranges exhibit considerable overlap. We require the same properties from their population counterparts to have good power. It should be noted at this point that the shape of the estimated propensity scores is already indicative of the way that miscarriage fails as an instrument. In a naive telling of the story, the propensity score for women who had a teenage miscarriage is shifted upward, contrary to what we observe in Figure 1.2. Our test rejects if, keeping the probability of treatment fixed, the difference between the outcomes of the subpopulation with $Z = 0$ and the subpopulation with $Z = 1$ is large. Figure 1.3 plots $\hat{m}_{0,j}(x) - \hat{m}_{1,j}(x)$ for all values of j. The dashed lines indicate our estimates of $x_{L,j}$ and $x_{U,j}$. We observe that the estimated outcome difference is positive and decreasing in the probability of treatment x. This means that for a low treatment probability x women who have a miscarriage do much worse in terms of high school graduation than do women who do not have a miscarriage. For larger x, however, this difference in outcomes becomes smaller. This feature is in line with our story-based criticism of the instrument. Suppose that the underlying heterogeneity selects women into pregnancy rather than into motherhood. For concreteness think of the heterogeneity as the amount of unprotected sex that a woman has and suppose that this variable is highly correlated with outcomes. In a Bayesian

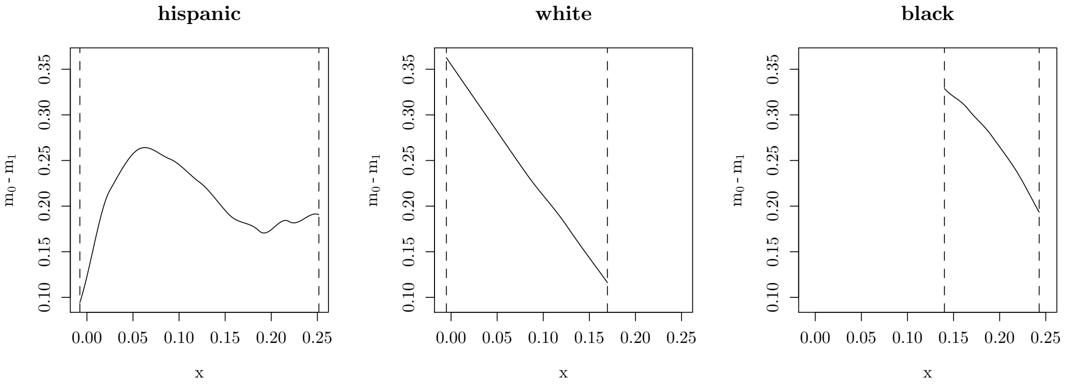


Figure 1.3: Difference in expected outcomes conditional on probability of treatment between the subpopulations with $Z = 0$ and $Z = 1$. Plotted with $q = 1$ and bandwidths $h = 0.25$ and $g = 2.00$.

sense a woman who has a miscarriage reveals herself to be of the type that is prone to have unprotected sex. In that sense she is very similar to women with a high probability of becoming pregnant and carrying the child to term and very different from women who become pregnant only with small probability. To turn this eye-balling of the plots in Figure 1.3 into a rigorous argument we now take into account sampling error by applying our formal testing procedure. For both the first and the second stage regression we choose an Epanechnikov kernel. To have good power against local alternatives we choose $q = 2$. To keep the problem tractable and to reduce the number of parameters we have to choose, we set $g_j = g$ and $h_j = h$ for all j. We then run the test for a large number of bandwidth choices letting h vary between 0.1 and 0.5 and letting g vary between 1 and 3. To determine the bounds of integration $\underline{x}_j$ and $\bar{x}_j$ we use the naive sample equivalence approach suggested in Section 1.4 with different values for c_δ. Table 1.5 in the appendix reports results for $c_\delta = 0.05$ and Table 1.6 reports results for $c_\delta = 0.075$. For these two choices of c_δ the test rejects at moderate to high significance levels for a large range of smoothing parameter choices.

Our approach can also be used to investigate other instruments that have been suggested in the literature on teen pregnancies. For example, Z or S could be based on local variation in abortion rates or in availability of fertility related health services (cf. Ribar 1994; Klepinger, Lundberg, and Plotnick 1995).

1.7 Conclusion

So far, inference about heterogeneous treatment effect models mostly relies on theoretical considerations about the relationship between instruments and unobserved individual characteristics that are not investigated empirically. This paper shows that under the assumption that a binary and a continuous instrument are available, a parameter is overidentified. This provides a way to test whether the model is correctly specified. The overidentification result is not merely a theoretical curiosity, it has bite when applied to real data. We illustrate this by applying our method to a dataset on teenage child bearing and high school graduation.

Apart from suggesting a new test, we also contribute to the statistical literature by developing testing theory that with slight modifications can be applied to other settings where index sufficiency holds under the null hypothesis. We accommodate an index that is not observed and enters the test statistic as a nonparametrically generated regressor. This setting is encountered, e.g., when testing the validity of the matching approach along the lines suggested in Heckman et al. 1996 and Heckman et al. 1998. Heckman et al. 1998 employ a parametric first-stage estimator. As a result, their second-stage estimator is, to first order, identical to the oracle estimator. Our analysis suggests that replacing the parametric first-stage estimator by a non- or semiparametric estimator is not innocuous. In particular, it can affect the second-stage bandwidth choice and the behavior of the test under local alternatives.

A theoretical analysis of our wild bootstrap procedure is beyond the scope of this paper. Developing resampling methods for models with nonparametrically generated regressors is an interesting direction for future research. We hope to corroborate the findings in our exploratory simulations by theoretical results in the future.

To apply our method to a particular data set, additional considerations might be necessary. In many applications the validity of an instrument is only plausible provided that a large set of observed variables is controlled for. It is hard to accommodate a rich covariate space in a completely nonparametric model. This is partly due to a curse of dimensionality. Another complicating factor is that our testing approach has good power only if, for fixed covariate values, the instruments provide considerable variation in participation. This is what allows us to test the model for a wide range of unobserved types. Typically, however, instruments become rather weak once the model is endowed with a rich covariate space. These issues can be dealt with by imposing a semiparametric model. As an example, consider the following simple variant of a model suggested in Carneiro and S. Lee 2009. We let X denote a vector of covariates with possibly continuous components and assume that the unobserved type V is independent of X. Treatment status is determined by $D = 1_{\{R \geq V\}}$ with $R = r_1(X) + r_2(S, Z)$. The unobserved type affects the treatment effect and not the base outcome. The observed outcome is

$$Y = \mu_\alpha(X) + D[\mu_\beta(X) + \lambda(V)].$$

The functions r_1, μ_α and μ_β are known up to a finite dimensional parameter. A semiparametric version of our test would compare $\mathrm{E}[D\lambda(V) \mid R = x, Z]$ in the $Z = 0$ and $Z = 1$ subpopulations. The fact that X is uninformative about V and the additive structure allow for an overidentification result that uses variation in X to extend the interval on which a function is overidentified. This contrasts sharply with Proposition 1.1 which relies on variation in S keeping the value of covariates fixed. In terms of asymptotic rates this semiparametric model with a large covariate space is not harder to estimate than our fully nonparametric model with a small covariate space and there is no curse of dimensionality.

As seen in Section 1.6 plots of the quantities underlying the test statistic can be helpful in interpreting test results and are a good starting point for discovering the source of a rejection. In many applications it is plausible to assume that while instruments are not valid for extreme types (types with a particularly low or high propensity to participate), they work well for the more average types. The plots can be used to heuristically identify the subpopulation for which instruments are valid. For a subpopulation that based on theoretical considerations is hypothesized to satisfy instrument validity, our approach offers a rigorous way of testing the correct specification of the subpopulation.

Appendix 1.A Definition of estimators

Let $L_g(\cdot) = g^{-1}L(\cdot/g)$ and $K_h(\cdot) = h^{-1}K(\cdot/h)$. For the first-stage estimator set $\hat{r}_{z,j}(s) = a_0$, where a_0 satisfies

$$(a_0, \dots, a_q) \in \arg\min_{(a_0,\dots,a_q)\in\mathbb{R}^{q+1}} \sum_{i:Z_i=z,J_i=j} \big(D_i - a_0 - a_1(S_i - s) - \cdots \\ - a_q(S_i - s)^q\big)^2 L_g(S_i - s).$$

For the second-stage estimator set $\hat{m}_{z,j}(x) = b_0$, where b_0 satisfies

$$(b_0, b_1) \in \arg\min_{(b_0,b_1)\in\mathbb{R}^2} \sum_{i:Z_i=z,J_i=j} \big(Y_i - b_0 - b_1(\hat{r}_{z,j}(S_i) - x)\big)^2 K_h(\hat{r}_{z,j}(S_i) - x).$$

Appendix 1.B Proofs

Proof of Theorem 1.2

The proposition follows from a sequence of lemmas. We first prove that the second-stage regression function and the error terms from the first- and second-stage regressions behave nicely under our assumptions about the primitives of the model.

Assumption 1.3
For each $z \in \{0, 1\}$

(i) m_z *is twice continuously differentiable on* (x_L, x_U).

(ii) there is a positive ρ *such that* $\mathrm{E}_z[\exp(\rho|\zeta|) \mid S]$ *and* $\mathrm{E}_z[\exp(\rho|\epsilon|) \mid S]$ *are bounded,*

(iii) $\sigma^2_{\zeta,z}(x) = \mathrm{E}_z[\zeta^2 \mid r_z(S) = x]$, $\sigma^2_{\epsilon,z}(x) = \mathrm{E}_z[\epsilon^2 \mid r_z(S) = x]$, *and* $\sigma_{\epsilon\zeta,z}(x) = \mathrm{E}_z[\epsilon\zeta \mid r_z(S) = x]$ *are continuous on* (x_L, x_U).

Lemma 1.1 *Assumption 1.1 is sufficient for Assumption 1.3.*

Proof The lemma follows from plugging in the structural treatment model into the observed quantities and arguing similarly to the proof of Theorem 1.1. □

In the next lemma we give a complete description of the relevant properties of our first-stage estimator. We provide an explicit expression of a smoothed version of the first-stage estimator that completely characterizes the impact of estimating the regressors on the asymptotic behavior of the test statistic.

Lemma 1.2 (First stage estimator) *The first stage local polynomial estimator can be written as*

$$\hat{r}_z(s) = \rho_n(s) + R_n,$$

where

$$\sup_s |R_n| = O_p\left(g^{q+1}\sqrt{\frac{\log n}{ng}} + \frac{\log n}{ng}\right)$$

and ρ_n *is given explicitly in equation* (1.8). *Wpa1* ρ_n *is contained in a function class* $\mathcal{R}$ *that for some constant* K*, any* $\xi > \frac{5}{4}\eta^* - \frac{1}{4}$ *and all* $\epsilon > 0$ *can be covered by* $K\exp(n^{\xi}\epsilon^{-1/2})$ ϵ*-balls with respect to the sup norm. The true propensity score is contained in* $\mathcal{R}$*. Furthermore,*

$$-m'(x)\int K_h(r_z(s)-x)(\hat{r}_z(s)-r_z(s))f_{S|Z=z}(s)\,ds = \frac{1}{n}\sum_{i:Z_i=z}\psi^{(2)}_{n,z,i}(x) + o_p\big(n^{-1/2}\big),$$

with $\psi^{(2)}_{n,z,i}$ *as defined in Lemma 1.3. Moreover,*

$$\sup_s |\hat{r}_z(s) - r_z(s)| = O_p\left(n^{-\frac{1}{2}(1-\eta^*)}\right).$$

Proof Throughout, condition on the subsample with $Z = z$. Let $e_1 = (1, 0, \dots, 0)^\top$ and $\mu(t) = (1, t, \dots, t^q)^\top$. Furthermore, define

$$\bar{M}_n(s) = \mathrm{E}\,\mu\left(\frac{S_i - s}{g}\right)\mu^\top\left(\frac{S_i - s}{g}\right)L_g(S_i - s).$$

Since we defined g in terms of the total sample size it behaves like a random variable when we work conditionally on the subsample $Z = z$. We have $g = a_n n_z^{-\eta^*} + O_p\left(n^{-\frac{1}{2}-\eta^*}\right)$ for a bounded deterministic sequence a_n. From a straightforward extension of standard arguments for the case of a deterministic bandwidth (c.f. Masry 1996) it can be shown that $\hat{r}_z$ can be written as

$$\hat{r}_z(s) = \rho_n(s) + R_n,$$

where

$$\rho_n(s) = r_z(s) + g^{q+1}b_n(s) + e_1^\top \bar{M}_n^{-1}(s)\frac{1}{n}\sum_i \mu\left(\frac{S_i - s}{g}\right)L_g(S_i - s)\zeta_i, \tag{1.8}$$

b_n is a bounded function and R_n has the desired order. To show that the desired entropy condition holds, note that $\bar{M}_n$ is a deterministic sequence that is bounded away from zero so that it suffices to derive an entropy bound for the functions

$$\frac{1}{n}\sum_i \mu\left(\frac{S_i - s}{g}\right)L_g(S_i - s)\zeta_i.$$

Wpa1 these functions have a second derivative that is bounded by $\sqrt{n^{-1}g^5\log n}$. The desired bound on the covering number then follows from a straightforward corollary to Theorem 2.7.1 in van der Vaart and Wellner 1996. To prove the statement about the smoothed first stage estimator note that under our assumptions we only have to consider the smoothed error term

$$\frac{1}{n}\sum_{i:Z_i=z}\psi_n^*(x,S_i)\zeta_i,$$

where

$$\begin{aligned}\psi_n^*(x,s) &= -m'(x)\int K_h(r_z(u)-x)e_1'\bar{M}_n^{-1}(u)\mu\left(\frac{s-u}{g}\right)L_g(s-u)f_{S|Z=z}(u)\,du\\ &= -m'(x)\int K_h(r_z(s-gu)-x)e_1'\bar{M}_n^{-1}(s-gu)\mu(u)L(u)f_{S|Z=z}(s-gu)\,du.\end{aligned}$$

Since $f_{S|Z=z}$ is bounded and has a bounded derivative there is a function $D_n(s,u)$ bounded uniformly in s, u and x satisfying

$$\bar{M}_n^{-1}(s-ug)f(s-ug)-M^{-1}=gD_n(s,u).$$

By standard kernel smoothing arguments

$$\frac{1}{n_z}\sum_{i:Z_i=z}\left\{\int K_h(r_z(S_i-ug)-x)D_n(S_i,u)\mu(u)L(u)\,du\right\}\zeta_i=O_p\left(\sqrt{\frac{\log n}{nh}}\right).$$

Noting that $L^*(u)=e_1^\top M^{-1}\mu(u)L(u)$ we have

$$\frac{1}{n}\sum_{i:Z_i=z}\psi_n^*(x,S_i)\zeta_i=\frac{1}{n}\sum_{i:Z_i=z}\psi_{n,z,i}^{(2)}(x)+o_p(n^{-1/2}).$$

□

Next, we give an asymptotic expansion of the integrand in (1.4) up to parametric order. The result states that the integrand can be characterized by a deterministic function that summarizes the deviation from index sufficiency under the alternative and an asymptotic influence function calculated under the hypothetical model $\mathcal{M}^{\text{null}}$.

Lemma 1.3 (Expansion) *Uniformly in x*

$$\hat{m}_0(x)-\hat{m}_1(x)=\Delta_{K,h}(x)+\frac{1}{n}\sum_i\psi_{n,i}(x)+o_p(n^{-1/2})$$

where $\psi_{n,i}=\psi_{n,i}^{(1)}+\psi_{n,i}^{(2)}$ and $\psi_{n,i}^{(j)}=\sum_{z=0,1}\psi_{n,z,i}^{(j)}$, $j=1,2$,

$$\begin{aligned}\psi_{n,z,i}^{(1)}(x) &= \frac{1_{\{Z_i=z\}}(-1)^z}{p_z f_{R,z}(x)}K_h(r_z(S_i)-x)\epsilon_i,\\ \psi_{n,z,i}^{(2)}(x) &= -m'(x)\frac{1_{\{Z_i=z\}}(-1)^z}{p_z f_{R,z}(x)}\int K_h(r_z(S_i-gu)-x)L^*(u)\,du\,\zeta_i.\end{aligned}$$

Here $\epsilon_i=Y^{null}-\mathrm{E}[Y^{null}\mid r_Z(S)]$, i.e, ϵ_i is the residual under the hypothetical model $\mathcal{M}^{null}$, and L^* denotes the equivalent kernel of the first step local polynomial regression.

Proof The statement follows from an expansion of $\hat{m}_z$. Work conditionally on the subsample with $Z = z$ and let n_z denote the number of observations in the subsample. To avoid confusion, we write h_n for the second-stage bandwidth, as h will sometimes denote a generic element of a set of bandwidths. Let $h^z = n_z^{-\eta}$. Note that for C large enough h_n is contained in the set

$$\mathcal{H}_{n_z} = \left\{h' : \left|h' - h^z\right| \leq C n_z^{-1/2-\eta}\right\}$$

wpa1. Let $e_1 = (1,0)^\top$, $\mu(t) = (1,t)^\top$ and

$$M_h^r(x) = \frac{1}{n}\sum_{i:Z_i=z} \mu\big((r(S_i)-x)/h\big)\mu^\top\big((r(S_i)-x)/h\big)K_h(r(S_i)-x).$$

For arbitrary $\mathbb{R}^n$-valued random variables W define the local linear smoothing operator

$$\mathcal{K}^r_{h,x,z}W = e_1^\top\left(M_h^r(x)\right)^{-1}\frac{1}{n_z}\sum_{i:Z_i=z} W_i\mu\left(\frac{r(S_i)-x}{h}\right)K_h(r(S_i)-x).$$

Decompose the estimator as

$$\begin{aligned}\hat{m}_z(x) =&\mathcal{K}^{\hat r}_{h_n,x,z}Y^n + \mathcal{K}^{\hat r}_{h_n,x,z}\left\{(Y^n - Y^{\text{null}}) - \mathrm{E}[Y^n - Y^{\text{null}} \mid S, Z]\right\}\\ &+ \mathcal{K}^{\hat r}_{h_n,x,z}\,\mathrm{E}[Y^n - Y^{\text{null}} \mid S, Z]\\ =&J_1 + J_2 + J_3.\end{aligned}$$

We now proceed to show that

$$\begin{aligned}&J_1 = m(x) + b_{1,n}(x) + \frac{1}{n}\sum_i\left\{\psi^{(1)}_{n,z,i}(x) + \psi^{(2)}_{n,z,i}(x)\right\} + o_p\big(n^{-1/2}\big),\\ &J_2 = o_p\big(n^{-1/2}\big),\\ &J_3 = b_{2,n}(x) + \int \mathrm{E}[\varphi_n(S,Z) \mid r_Z(S) = x + hr, Z = z]K(r)\,dr + o_p\big(n^{-1/2}\big),\end{aligned}$$

where $b_{j,n}$, $j = 1,2$, are independent of z and all order symbols hold uniformly in x. For the J_1 term we apply the approach from Mammen, Rothe, and Schienle 2012 (MRS) and expand J_1 around the oracle estimator. Write

$$J_1 = \mathcal{K}^{\hat r}_{h_n,x,z}\epsilon_i + \mathcal{K}^{\hat r}_{h_n,x,z}m(r_z(S_i)) = J_{1,a} + J_{1,b}.$$

For the $J_{1,a}$ term note that $e_1^\top\left(M_h^r(x)\right)^{-1}$ is stochastically bounded by a uniform over $\mathcal{H}_{n_z}$ version of Lemma 2 in MRS. For ρ_n as defined in Lemma 1.2 write

$$\begin{aligned}&\frac{1}{n_z}\sum_{i:Z_i=z}K_{h_n}(\hat r_z(S_i) - x)\epsilon_i - \frac{1}{n_z}\sum_{i:Z_i=z}K_{h_n}(r_z(S_i) - x)\epsilon_i\\ =&\frac{1}{n_z}\sum_{i:Z_i=z}\left(K_{h_n}(\hat r(S_i) - x) - K_{h_n}(\rho_n(S_i) - x)\right)\epsilon_i\\ &+\frac{1}{n_z}\sum_{i:Z_i=z}\left(K_{h_n}(\rho_n(S_i) - x) - K_{h_n}(r_z(S_i) - x)\right)\epsilon_i = I_1 + I_2.\end{aligned}$$

By the mean-value theorem $I_1 = o_p(n^{-1/2})$. For I_2 note that $\mathrm{E}_z[\epsilon \mid S] = 0$ so that following the arguments in the proof of Lemma 2 in MRS

$$\sup_{x;h\in\mathcal{H}_{n_z}} P\left(\sup_{r_1,r_2\in\mathcal{R}}\left|\frac{1}{n_z}\sum_{i:Z_i=z}\left(K_h(r_1(S_i)-x)-K_h(r_2(S_i)-x)\right)\epsilon_i\right| > C^* n^{-\kappa_1}\right) \\ \le \exp(-cn^c),$$

where κ_1 is defined in MRS and C^* is a large constant. To check that $\kappa_1 > 1/2$ note that Theorem 1 in MRS allows bandwidth exponents in an open set so that it suffices to check the conditions for h^z. It is now straightforward to show that a polynomial number of points in $[\underline{x}, \bar{x}] \times \mathcal{H}_{n_z}$ provide a good enough approximation to ensure that

$$\sup_{x,h\in\mathcal{H}_{n_z},\rho\in\mathcal{R}}\left|\frac{1}{n_z}\sum_{i:Z_i=z}\left(K_h(\rho(S_i)-x)-K_h(r_z(S_i)-x)\right)\epsilon_i\right| = O_p(n^{-\kappa_1})$$

and hence $I_2 = o_p(n^{-1/2})$. Similar arguments apply to

$$\frac{1}{n_z}\sum_{i:Z_i=z}\frac{\hat{r}_z(S_i)-x}{h_n}K_{h_n}(\hat{r}_z(S_i)-x)\epsilon_i.$$

Therefore, $J_{1,a}$ can be replaced by its oracle counterpart at the expense of a remainder term that vanishes at the parametric rate:

$$J_{1,a} = \frac{1}{n}\sum_i \psi^{(1)}_{n,z,i}(x) + o_p(n^{-1/2}).$$

Note that in the last step we also replaced n_z by $p_z n$. Decompose $J_{1,b}$ as in the proof of Theorem 1 in MRS. It is straightforward to extend their results to hold uniformly over bandwidths in $\mathcal{H}_{n_z}$. Deduce that

$$J_{1,b} = m(x) + b_{1,n}(x) - m'(x)\int K_{h_n}(r_z(s)-x)(\hat{r}_z(s)-r_z(s))f_{S|Z=z}(s)\,ds + o_p\big(n^{-1/2}\big).$$

for a sequence of functions $b_{1,n}$ that does not depend on the design. The previous results use standard results about the Bahadur representation of the oracle estimator (cf. Masry 1996; Kong, Linton, and Xia 2010). The desired representation for J_1 follows from Lemma 1.2. For the J_2 term apply Lemma 2 in MRS in a similar way as described above to argue that

$$J_2 - \mathcal{K}^{r_z}_{h_n,x,z}\left\{(Y^n - Y^{\mathrm{null}}) - \mathrm{E}[Y^n - Y^{\mathrm{null}} \mid S, Z]\right\} = J_2 - J_2^* = o_p\big(n^{-1/2}\big).$$

By standard kernel smoothing arguments $J_2^* = o_p\big(n^{-1/2}\big)$. For the J_3 term let $A_i = \mathrm{E}[Y_i^n - Y_i^{\mathrm{null}} \mid S_i, Z_i]$ and consider the behavior of the terms

$$\frac{1}{n_z}\sum_{i:Z_i=z}A_i\left(\frac{\hat{r}_z(S_i)-x}{h_n}\right)^a K_{h_n}(\hat{r}_z(S_i)-x), \quad a = 0, 1.$$

We focus on $a = 0$. The argument for the other case is similar. Let $K'_h(\cdot) = h^{-1}K'(\cdot/h)$. For any $\tilde{r}$ (pointwise) between $\hat{r}_z$ and r_z

$$\sup_x \left| \frac{1}{n_z} \sum_{i:Z_i=z} K'_{h_n}(\tilde{r}(S_i) - x) \right| \leq C \sup_x \frac{1}{n_z h^z} \sum_{i:Z_i=z} 1_{\{|r_z(S_i)-x| \leq Ch^z\}} = O_p(1)$$

for a positive constant C. Noting that $\max_{i \leq n} |A_i| = O_p(c_n)$ it is now easy to see that

$$\begin{aligned} &\frac{1}{n_z} \sum_{i:Z_i=z} A_i K_{h_n}(\hat{r}_z(S_i) - x) \\ =&\frac{1}{n_z} \sum_{i:Z_i=z} A_i \left[K_{h_n}(r_z(S_i) - x) + K'_{h_n}(\tilde{r}(S_i) - x) \frac{\hat{r}_z(S_i) - r_z(S_i)}{h_n} \right] \\ =&\frac{1}{n_z} \sum_{i:Z_i=z} A_i K_{h_n}(r_z(S_i) - x) + o_p\big(n^{-1/2}\big) \end{aligned}$$

uniformly in x. Let $M = \int \mu(t)\mu^\intercal(t)K(t)\,dt$, $M_n = M^{r_z}_{h_n}$ and $\bar{M}_n = \mathrm{E}\, M_n$. By Lemma 2 in Mammen, Rothe, and Schienle 2012 and standard arguments we have

$$\begin{aligned} M^{\hat{r}_z}_{h_n}(x) - f_{R|Z=z}M =&M^{\hat{r}_z}_{h_n}(x) - M_n(x) + M_n(x) - \bar{M}_n(x) + \bar{M}_n(x) - f_{R|Z=z}(x)M \\ =&O_p\left(n^{-\frac{1}{2}(1-3\eta)} + \sqrt{\frac{\log n}{nh_n}} + h_n \right) \end{aligned}$$

uniformly in x. Therefore,

$$J_3 - f^{-1}_{R|Z=z}(x) \frac{1}{n_z} \sum_{i:Z_i=z} A_i K_{h_n}(r_z(S_i) - x) = J_3 + J_3^* = o_p\big(n^{-1/2}\big).$$

It is straightforward to show that under our assumptions J_3^* can be replaced by its expectation at the expense of an uniform $o_p(n^{-1/2})$ term. Since

$$\mathrm{E}[Y^n - Y^{\text{null}} \mid S, Z] = \mathrm{E}[Y^n - Y^{\text{null}} \mid r_Z(S)] + \varphi_n(S, Z),$$

and since $f_{R|Z=z}$ has a bounded derivative

$$\begin{aligned} \mathrm{E}_z\, J_3^* =& \int \mathrm{E}[Y^n - Y^{\text{null}} \mid r_Z(S) = x + h_n r] K(r)\,dr \\ &+ \int \mathrm{E}[\varphi_n(S, Z) \mid r_Z(S) = x + h_n r, Z = z] K(r)\,dr + o\big(n^{-1/2}\big). \end{aligned}$$

Here we keep implicit that we are treating h_n as a constant in the above expectations, i.e., we are integrating with respect to the marginal measure of (Z, S). The conclusion follows by noting that the first term on the right-hand side is independent of z. □

Plugging in from Lemma 1.3 gives an asymptotic expansion of the test statistic.

Lemma 1.4

$$T_n = T_{n,a} + T_{n,b} + \int \Delta_{K,h}^2(x)\,dx + o_p(n\sqrt{h}),$$

where

$$T_{n,a} = \frac{2}{n^2}\sum_{i<j}\int \psi_{n,i}(x)\psi_{n,j}(x)\,dx \quad \text{and} \quad T_{n,b} = \frac{1}{n^2}\int \sum_i \psi_{n,i}^2(x)\,dx.$$

Proof Plug in from Lemma 1.3, expand the square and inspect each term separately.□

Lemma 1.5 (Variance) *For $T_{n,a}$ as defined in Lemma 1.4*

$$\begin{aligned} &\operatorname{var}(T_{n,a}) = n^{-2}h^{-1}V + o\left(n^{-2}h^{-1}\right) \quad \text{and} \\ &n\sqrt{h}T_{n,a} \xrightarrow{d} \mathcal{N}(0,V). \end{aligned}$$

Proof For the first part of the lemma, note that

$$\int K_h(r_z(s-gu)-x)L^*(u)\,du = \int \Big\{K_h(r_z(s)-x) + \underbrace{K'(\chi_1/h)\partial_s r_z(\chi_2)u}_{\equiv a(s,u,x)}\frac{g}{h^2}\Big\}L^*(u)\,du,$$

where χ_1 is an intermediate value between $r_z(s-hu)-x$ and $r_z(s)-x$, and χ_2 is an intermediate value between $s-hu$ and s. As K and r_z have bounded derivatives

$$\tilde{a}(r,x) = \mathrm{E}\Big[\int a(S,u,x)L^*(u)\,du \mid r_z(S) = r\Big]$$

is a bounded function. By standard U-statistic arguments

$$\begin{aligned} &\operatorname{var}\left(2\sum_{i<j}\int \psi_{n,i}(x)\psi_{n,j}(x)\,dx\right) = 4\sum_{i<j}\mathrm{E}\Big[\int \psi_{n,i}(x)\psi_{n,j}(x)\,dx\Big]^2 \\ &= 4\binom{n}{2}\int h\left\{\mathrm{E}[\psi_{n,1}(x)\psi_{n,1}(x+hx')]\right\}^2\,dx'\,dx. \end{aligned}$$

Note that

$$\mathrm{E}[\psi_{n,1}(x)\psi_{n,1}(x+hx')] = \sum_{z\in\{0,1\}}\mathrm{E}[\psi_{n,z,1}(x)\psi_{n,z,1}(x+hx')].$$

We consider here only one of the terms composing $\mathrm{E}[\psi_{n,z,1}(x)\psi_{n,z,1}(x+hx')]$. For the other terms similar arguments apply. Let

$$q(x) = -\frac{m'(x)1_{\{Z=z\}}}{p_z f_{R|Z=z}}\int K_h(r_z(S-gu)-x)L^*(u)\,du.$$

Using $\mathrm{E}_z[\zeta^2 \mid r_Z(S) = x] = x(1-x)$ we have

$$\begin{aligned}
&h[\mathrm{E}\, q(x)q(x+hx')\zeta_1^2] \\
=&h\,\mathrm{E}\left[\frac{1_{\{Z=z\}}m_z'(x)m_z'(x+hx')}{p_z^2 f_{R|Z=z}(x) f_{R|Z=z}(x+hx')}(K_h(r_z(S)-x)+\frac{g}{h^2}\tilde{a}(r_z(S),x))\cdots\right. \\
&\left.\cdots(K_h(r_z(S)-x-hx')+\frac{g}{h^2}\tilde{a}(r_z(S),x-hx')\zeta^2\right] \\
=&\frac{x(1-x)[m_z'(x)]^2}{p_z f_{R|Z=z}(x)}\int K(y)K(x'-y)\,dy + o(1) = \frac{x(1-x)[m_z'(x)]^2}{p_z f_{R|Z=z}(x)}K^{(2)}(x') + o(1).
\end{aligned}$$

For the second part of the lemma it suffices to check the two conditions of Theorem 2.1 in de Jong 1987. Let $W_{ij} = 2n^{-1}\sqrt{h}\int \psi_i(x)\psi_j(x)$ and show that

$$\mathrm{var}^{-1}\Big(\sum_{i<j} W_{ij}\Big)\max_{1\le i\le n}\sum_{1\le j\le n}\mathrm{var}(W_{ij}) \to 0$$

$$\mathrm{var}^{-2}\Big(\sum_{i<j} W_{ij}\Big)\,\mathrm{E}\Big\{\sum_{i<j} W_{ij}\Big\}^4 \to 3.$$

The first condition holds trivially. To show that the second condition is satisfied note that $\mathrm{var}(\sum_{i<j} W_{ij})$ converges to a constant. It is easy to see that asymptotically only terms of the form $\mathrm{E}\,W_{ij}^2 W_{kl}^2$ with $\{i,j\}\cap\{k,l\} = \varnothing$ will contribute to $\mathrm{E}\left[\sum_{i<j} W_{ij}\right]^4$. There are

$$\binom{4}{2}\frac{\binom{n}{2}\left[\binom{n}{2}-1\right]}{2!} \approx \frac{3}{4}n^4$$

such terms when expanding $\mathrm{E}\left[\sum_{i<j} W_{ij}\right]^4$. The condition then follows by noting that

$$\mathrm{var}\Big(\sum_{i<j} W_{ij}\Big) = \sum_{i<j}\mathrm{E}\,W_{ij}^2$$

and that $\mathrm{E}\,W_{ij}^2 W_{kl}^2$ factors as $\mathrm{E}\,W_{ij}^2\,\mathrm{E}\,W_{kl}^2$. □

We now apply standard U-statistic theory. As the next two lemmas show, $T_{n,b}$ contributes to the asymptotic bias and $T_{n,b}$ contributes to the asymptotic variance.

Lemma 1.6 (Bias) *For $T_{n,b}$ as defined in Lemma 1.4*

$$n\sqrt{h}T_{n,b} = \frac{1}{\sqrt{h}}\gamma_n + o_p(1),$$

where γ_n is a deterministic sequence converging to γ.

Proof Write

$$n\sqrt{h}T_{n,b} = \frac{\sqrt{h}}{n}\sum_i\int\psi_{n,i}^2(x)\,dx = \mathrm{E}\left\{\frac{\sqrt{h}}{n}\sum_i\int\psi_{n,i}^2(x)\,dx\right\} + o_p(1) \equiv \gamma_n + o_p(1).$$

Define the function a as in the proof for Lemma 1.5. To compute γ_n write

$$\begin{aligned}\psi^2_{n,z,i}(x) =&\frac{1_{\{Z=z\}}}{p_z^2 f_{R,z}^2(x)}\Big\{K_h^2(r_z(S)-x)\epsilon^2 + [m'(x)]^2 K_h^2(r_z(S)-x)\zeta^2\\ &-2m'(x)K_h(r_z(S)-x)\epsilon\zeta\Big\}\\ &+\frac{1_{\{Z=z\}}}{p_z^2 f_{R,z}^2(x)}\left(\frac{g}{h^2}\int g(S,u,x)L^*(u)\,du\right)^2\zeta^2+\\ &\frac{1_{\{Z=z\}}}{p_z^2 f_{R,z}^2(x)}\frac{g}{h^2}\left(\int g(S,u,x)L^*(u)\,du\right)K_h(r_z(S)-x)\epsilon\zeta\\ =&\Gamma_1(S,x)+\Gamma_2(S,x)+\Gamma_3(S,x).\end{aligned}$$

Note that

$$h\sum_{z=0,1}\mathrm{E}\int\Gamma_1(S,x)\,dx\to\gamma,$$

where we kept the dependence of Γ_1 on z implicit. Now show that the other terms entering γ_n vanish. To show that $h\sum_{z=0,1}\mathrm{E}\int\Gamma_3(S,x)\,dx\to 0$ it suffices to show that

$$\mathrm{E}_z\left[\left(\int g(S,u,x)L^*(u)\,du\right)\epsilon\zeta \mid r_z(S)\right]$$

is bounded. This follows immediately from the fact that $\int g(S,u,x)L^*(u)\,du$ is bounded and hence

$$\mathrm{E}_z\left[\int g(S,u,x)L^*(u)\,du\ \epsilon\zeta \mid r_z(S)\right]\lesssim \mathrm{E}_z[|\epsilon\zeta| \mid r_z(S)]\le\sqrt{\sigma_\epsilon^2(r_z(S))}\le C$$

for some constant C. For $h\sum_{z=0,1}\mathrm{E}\int\Gamma_2(S,x)\,dx$ argue similarly. □

Proof of Theorem 1.3

Proof Using the expansion from Lemma 1.3 and applying standard smoothing arguments to the stochastic term we get that for a small enough open set $\mathcal{G}_x\supset[\underline{x},\bar{x}]$

$$\sup_{x\in\mathcal{G}_x}|\hat{m}_0(x)-\hat{m}_1(x)|^2 = O\left(\frac{1}{n\sqrt{h}}+g^{2(q+1)}\right)+O_p\left(\frac{\log n}{nh}\right)+o_p\left(\frac{1}{n}\right).$$

Write

$$T_n(\underline{x}_n,\bar{x}_n)-T_n(\underline{x},\bar{x}) = T_n(\underline{x}_n,\underline{x})-T_n(\bar{x},\bar{x}_n).$$

We can bound $T_n(\underline{x}_n,\underline{x})$ by

$$|\underline{x}_n-\underline{x}|\sup_{x\in\mathcal{G}_x}|\hat{m}_0(x)-\hat{m}_1(x)| = o_p(n\sqrt{h}).$$

Similarly, we can find a bound for $T_n(\bar{x},\bar{x}_n)$. □

Appendix 1.C Tables

	$\theta = 0.10$						$\theta = 0.05$					
C_h	0.50	0.75	1.00	1.25	1.50	1.75	0.50	0.75	1.00	1.25	1.50	1.75
null												
$C_g = 0.50$	6.7	5.7	5.8	8.9	7.0	7.9	2.7	2.6	1.9	4.6	3.3	2.8
$C_g = 0.75$	9.2	6.4	8.2	6.5	6.4	7.0	4.6	2.0	3.2	2.8	3.2	2.8
$C_g = 1.00$	6.4	6.7	8.1	6.8	8.8	7.1	2.2	2.9	2.9	3.1	3.2	2.8
alternative 1												
$C_g = 0.50$	65.8	65.8	67.7	63.8	65.3	65.7	50.5	49.9	53.2	47.5	50.6	50.8
$C_g = 0.75$	65.1	65.8	64.8	65.3	65.8	65.9	49.7	47.7	49.9	49.5	50.1	52.3
$C_g = 1.00$	66.3	65.0	66.4	67.9	64.8	66.5	50.4	51.2	50.3	51.1	50.9	49.2
alternative 2												
$C_g = 0.50$	82.4	79.9	80.2	80.5	81.6	78.0	67.9	66.8	68.4	67.3	68.6	65.5
$C_g = 0.75$	79.2	81.0	79.9	80.6	80.4	79.8	66.1	68.3	68.0	68.2	66.5	65.8
$C_g = 1.00$	80.9	81.4	80.1	80.3	80.3	78.2	68.4	67.5	66.1	66.9	64.0	64.7
alternative 3												
$C_g = 0.50$	6.9	8.1	8.8	7.7	5.0	6.7	2.3	3.9	3.3	4.2	1.8	3.2
$C_g = 0.75$	7.2	8.1	6.8	7.4	6.7	6.9	2.9	2.6	3.7	3.9	2.1	3.0
$C_g = 1.00$	7.7	8.0	6.2	6.7	7.8	7.1	2.6	3.3	3.1	2.3	3.5	2.6
alternative 4												
$C_g = 0.50$	15.0	10.5	15.1	14.0	13.1	12.2	7.0	4.8	6.5	5.7	7.0	6.6
$C_g = 0.75$	12.5	13.9	13.8	12.9	13.6	13.3	5.2	6.2	7.0	7.0	6.3	5.9
$C_g = 1.00$	10.0	15.7	14.1	15.7	11.5	14.2	4.2	6.9	7.4	9.5	4.7	6.7
alternative 5												
$C_g = 0.50$	12.0	12.4	15.5	13.5	14.2	13.2	5.7	4.6	7.4	4.9	5.8	6.0
$C_g = 0.75$	13.4	14.5	12.5	12.1	12.0	11.1	6.0	6.9	5.7	4.2	5.3	5.7
$C_g = 1.00$	12.2	14.3	13.1	12.6	12.9	12.2	5.3	5.8	5.7	5.9	6.4	6.2
alternative 6												
$C_g = 0.50$	22.5	23.0	22.9	24.0	21.6	23.1	12.3	12.4	11.2	14.3	10.7	12.8
$C_g = 0.75$	23.3	20.6	25.3	23.5	23.3	20.0	12.5	11.4	13.2	12.0	13.5	12.1
$C_g = 1.00$	22.0	22.0	20.9	25.7	24.0	20.8	11.7	11.6	9.9	13.4	12.7	9.9

Table 1.3: Simulation. Empirical rejection probabilities in percentage points under nominal level θ. Sample size is $n = 200$.

			D		Y	
Race	Z	n	mean	sd	mean	sd
black	0	787	0.19949	0.3999	0.8183	0.3858
	1	67	0.26866	0.4466	0.6269	0.4873
hispanic	0	549	0.18033	0.3848	0.7687	0.4221
	1	36	0.27778	0.4543	0.5278	0.5063
white	0	1394	0.07389	0.2617	0.8479	0.3592
	1	77	0.20779	0.4084	0.6234	0.4877

Table 1.4: Teenage child bearing (D) and high-school graduation (Y).

	g	h	T_n	$\underline{x}_1$	$\bar{x}_1$	$\underline{x}_2$	$\bar{x}_2$	$\underline{x}_3$	$\bar{x}_3$	$P(>T_n)$	test result
1	1.00	0.15	0.086	0.03	0.32	0.01	0.21	0.05	0.46	0.225	no rejection
2	1.50	0.15	0.053	0.03	0.27	0.03	0.18	0.05	0.41	0.224	no rejection
3	2.00	0.15	0.084	0.03	0.21	0.02	0.15	0.05	0.36	0.059	*
4	2.50	0.15	0.054	0.04	0.19	0.02	0.13	0.11	0.27	0.012	**
5	3.00	0.15	0.022	0.02	0.18	0.05	0.11	0.13	0.19	0.092	*
6	1.00	0.20	0.064	0.03	0.32	0.01	0.21	0.05	0.46	0.060	*
7	1.50	0.20	0.042	0.03	0.27	0.03	0.18	0.05	0.41	0.084	*
8	2.00	0.20	0.067	0.03	0.21	0.02	0.15	0.05	0.36	0.010	**
9	2.50	0.20	0.043	0.04	0.19	0.02	0.13	0.11	0.27	0.010	**
10	3.00	0.20	0.019	0.02	0.18	0.05	0.11	0.13	0.19	0.083	*
11	1.00	0.25	0.045	0.03	0.32	0.01	0.21	0.05	0.46	0.012	**
12	1.50	0.25	0.037	0.03	0.27	0.03	0.18	0.05	0.41	0.036	**
13	2.00	0.25	0.051	0.03	0.21	0.02	0.15	0.05	0.36	0.008	***
14	2.50	0.25	0.035	0.04	0.19	0.02	0.13	0.11	0.27	0.025	**
15	3.00	0.25	0.017	0.02	0.18	0.05	0.11	0.13	0.19	0.090	*
16	1.00	0.30	0.040	0.03	0.32	0.01	0.21	0.05	0.46	0.010	**
17	1.50	0.30	0.035	0.03	0.27	0.03	0.18	0.05	0.41	0.036	**
18	2.00	0.30	0.044	0.03	0.21	0.02	0.15	0.05	0.36	0.021	**
19	2.50	0.30	0.030	0.04	0.19	0.02	0.13	0.11	0.27	0.022	**
20	3.00	0.30	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.080	*
21	1.00	0.35	0.039	0.03	0.32	0.01	0.21	0.05	0.46	0.005	***
22	1.50	0.35	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.024	**
23	2.00	0.35	0.041	0.03	0.21	0.02	0.15	0.05	0.36	0.014	**
24	2.50	0.35	0.029	0.04	0.19	0.02	0.13	0.11	0.27	0.018	**
25	3.00	0.35	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.064	*
26	1.00	0.40	0.038	0.03	0.32	0.01	0.21	0.05	0.46	0.003	***
27	1.50	0.40	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.021	**
28	2.00	0.40	0.040	0.03	0.21	0.02	0.15	0.05	0.36	0.007	***
29	2.50	0.40	0.028	0.04	0.19	0.02	0.13	0.11	0.27	0.011	**
30	3.00	0.40	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.064	*
31	1.00	0.50	0.038	0.03	0.32	0.01	0.21	0.05	0.46	0.003	***
32	1.50	0.50	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.012	**
33	2.00	0.50	0.040	0.03	0.21	0.02	0.15	0.05	0.36	0.012	**
34	2.50	0.50	0.029	0.04	0.19	0.02	0.13	0.11	0.27	0.005	***
35	3.00	0.50	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.065	*

Table 1.5: Test results for varying bandwidths and $c_\delta = 0.050$. (*) reject at 0.10 level, (**) reject at 0.05 level, (***) reject at 0.01 level.

	g	h	T_n	$\underline{x}_1$	$\bar{x}_1$	$\underline{x}_2$	$\bar{x}_2$	$\underline{x}_3$	$\bar{x}_3$	$P(> T_n)$	test result
1	1.00	0.15	0.057	0.06	0.29	0.03	0.18	0.07	0.44	0.170	no rejection
2	1.50	0.15	0.033	0.06	0.24	0.05	0.15	0.08	0.39	0.208	no rejection
3	2.00	0.15	0.066	0.06	0.19	0.04	0.13	0.07	0.33	0.042	**
4	2.50	0.15	0.037	0.06	0.16	0.04	0.10	0.13	0.25	0.011	**
5	3.00	0.15	0.009	0.04	0.16	0.07	0.09	0.16	0.16	0.114	no rejection
6	1.00	0.20	0.041	0.06	0.29	0.03	0.18	0.07	0.44	0.038	**
7	1.50	0.20	0.028	0.06	0.24	0.05	0.15	0.08	0.39	0.108	no rejection
8	2.00	0.20	0.048	0.06	0.19	0.04	0.13	0.07	0.33	0.009	***
9	2.50	0.20	0.029	0.06	0.16	0.04	0.10	0.13	0.25	0.014	**
10	3.00	0.20	0.009	0.04	0.16	0.07	0.09	0.16	0.16	0.101	no rejection
11	1.00	0.25	0.033	0.06	0.29	0.03	0.18	0.07	0.44	0.011	**
12	1.50	0.25	0.025	0.06	0.24	0.05	0.15	0.08	0.39	0.044	**
13	2.00	0.25	0.034	0.06	0.19	0.04	0.13	0.07	0.33	0.012	**
14	2.50	0.25	0.023	0.06	0.16	0.04	0.10	0.13	0.25	0.020	**
15	3.00	0.25	0.008	0.04	0.16	0.07	0.09	0.16	0.16	0.113	no rejection
16	1.00	0.30	0.031	0.06	0.29	0.03	0.18	0.07	0.44	0.010	**
17	1.50	0.30	0.024	0.06	0.24	0.05	0.15	0.08	0.39	0.036	**
18	2.00	0.30	0.031	0.06	0.19	0.04	0.13	0.07	0.33	0.015	**
19	2.50	0.30	0.020	0.06	0.16	0.04	0.10	0.13	0.25	0.023	**
20	3.00	0.30	0.007	0.04	0.16	0.07	0.09	0.16	0.16	0.138	no rejection
21	1.00	0.35	0.030	0.06	0.29	0.03	0.18	0.07	0.44	0.005	***
22	1.50	0.35	0.024	0.06	0.24	0.05	0.15	0.08	0.39	0.024	**
23	2.00	0.35	0.029	0.06	0.19	0.04	0.13	0.07	0.33	0.017	**
24	2.50	0.35	0.018	0.06	0.16	0.04	0.10	0.13	0.25	0.013	**
25	3.00	0.35	0.007	0.04	0.16	0.07	0.09	0.16	0.16	0.124	no rejection
26	1.00	0.40	0.030	0.06	0.29	0.03	0.18	0.07	0.44	0.008	***
27	1.50	0.40	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.020	**
28	2.00	0.40	0.029	0.06	0.19	0.04	0.13	0.07	0.33	0.016	**
29	2.50	0.40	0.028	0.04	0.19	0.02	0.13	0.11	0.27	0.005	***
30	3.00	0.40	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.076	*
31	1.00	0.50	0.038	0.03	0.32	0.01	0.21	0.05	0.46	0.001	***
32	1.50	0.50	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.012	**
33	2.00	0.50	0.040	0.03	0.21	0.02	0.15	0.05	0.36	0.012	**
34	2.50	0.50	0.029	0.04	0.19	0.02	0.13	0.11	0.27	0.010	**
35	3.00	0.50	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.062	*

Table 1.6: Test results for varying bandwidths and $c_\delta = 0.075$. (*) reject at 0.10 level, (**) reject at 0.05 level, (***) reject at 0.01 level.

References

Abadie, Alberto (2003). “Semiparametric instrumental variable estimation of treatment response models”. In: *Journal of Econometrics* 113.2, pp. 231–263.

Angrist, Joshua, Guido Imbens, and Donald Rubin (1996). “Identification of Causal Effects Using Instrumental Variables”. In: *Journal of the American Statistical Association* 91.434, pp. 444–455.

Balke, Alexander and Judea Pearl (1997). “Bounds on treatment effects from studies with imperfect compliance”. In: *Journal of the American Statistical Association* 92.439, pp. 1171–1176.

Carneiro, Pedro, James Heckman, and Edward Vytlacil (2011). “Estimating Marginal Returns to Education”. In: *American Economic Review* 101.6, pp. 2754–2781.

Carneiro, Pedro and Sokbae Lee (2009). “Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality”. In: *Journal of Econometrics* 149.2, pp. 191–208.

de Jong, Peter (1987). “A central limit theorem for generalized quadratic forms”. In: *Probability Theory and Related Fields* 75.2, pp. 261–277.

Delgado, Miguel A (1993). “Testing the equality of nonparametric regression curves”. In: *Statistics & probability letters* 17.3, pp. 199–204.

Dette, Holger and Natalie Neumeyer (2001). “Nonparametric analysis of covariance”. In: *the Annals of Statistics* 29.5, pp. 1361–1400.

Fernández-Val, Iván and Josh Angrist (2013). “ExtrapoLATE-ing: External validity and overidentification in the LATE framework”. In: Tenth World Congress. Advances in Economics and Econometrics: Theory and Applications 3. Econometric Society Monographs.

Frölich, Markus (2007). “Nonparametric IV estimation of local average treatment effects with covariates”. In: *Journal of Econometrics* 139.1, pp. 35–75.

Gørgens, Tue (2002). “Nonparametric comparison of regression curves by local linear fitting”. In: *Statistics & probability letters* 60.1, pp. 81–89.

Hall, Peter and Jeffrey D Hart (1990). “Bootstrap test for difference between means in nonparametric regression”. In: *Journal of the American Statistical Association* 85.412, pp. 1039–1049.

Hall, Peter and Joel Horowitz (2012). *A simple bootstrap method for constructing nonparametric confidence bands for functions.* Tech. rep. working paper.

Hansen, Lars Peter (1982). “Large sample properties of generalized method of moments estimators”. In: *Econometrica*, pp. 1029–1054.

Härdle, Wolfgang and Enno Mammen (1993). “Comparing nonparametric versus parametric regression fits”. In: *The Annals of Statistics* 21.4, pp. 1926–1947.

Heckman, James, Daniel Schmierer, and Sergio Urzua (2010). “Testing the correlated random coefficient model”. In: *Journal of Econometrics* 158.2, pp. 177–203.

Heckman, James, Sergio Urzua, and Edward Vytlacil (2006). “Understanding instrumental variables in models with essential heterogeneity”. In: *The Review of Economics and Statistics* 88.3, pp. 389–432.

Heckman, James and Edward Vytlacil (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation". In: *Econometrica*, pp. 669–738.

Heckman, James et al. (1996). "Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method". In: *Proceedings of the National Academy of Sciences* 93.23, pp. 13416–13420.

— (1998). "Characterizing selection bias using experimental data". In: *Econometrica: Journal of the Econometric Society* 66.5, pp. 1017–1098.

Hoffman, Saul D (1998). "Teenage childbearing is not so bad after all... or is it? A review of the new literature". In: *Family Planning Perspectives* 30.5, pp. 236–243.

Hotz, V Joseph, Susan Williams McElroy, and Seth G Sanders (2005). "Teenage Childbearing and Its Life Cycle Consequences Exploiting a Natural Experiment". In: *Journal of Human Resources* 40.3, pp. 683–715.

Hotz, V Joseph, Charles H Mullin, and Seth G Sanders (1997). "Bounding causal effects using data from a contaminated natural experiment: analysing the effects of teenage childbearing". In: *The Review of Economic Studies* 64.4, pp. 575–603.

Huber, Martin and Giovanni Mellace (2014). "Testing instrument validity for LATE identification based on inequality moment constraints". In: *Review of Economics and Statistics.*

Imbens, Guido and Joshua Angrist (1994). "Identification and estimation of local average treatment effects". In: *Econometrica*, pp. 467–475.

King, Eileen, Jeffrey D Hart, and Thomas E Wehrly (1991). "Testing the equality of two regression curves using linear smoothers". In: *Statistics & Probability Letters* 12.3, pp. 239–247.

Kitagawa, Toru (2013). "A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model". Working Paper.

Klepinger, Daniel H, Shelly Lundberg, and Robert D Plotnick (1995). "Adolescent fertility and the educational attainment of young women". In: *Family planning perspectives*, pp. 23–28.

Kong, Efang, Oliver Linton, and Yingcun Xia (2010). "Uniform bahadur representation for local polynomial estimates of M-regression and its application to the additive model". In: *Econometric Theory* 26.05, pp. 1529–1564.

Lee, Ying-Ying (2013). "Partial mean processes with generated regressors: Continuous Treatment Effects and Nonseparable models." Working Paper.

Levine, David I and Gary Painter (2003). "The schooling costs of teenage out-of-wedlock childbearing: analysis with a within-school propensity-score-matching estimator". In: *Review of Economics and Statistics* 85.4, pp. 884–900.

Mammen, Enno (1993). "Bootstrap and wild bootstrap for high dimensional linear models". In: *The Annals of Statistics*, pp. 255–285.

Mammen, Enno, Christoph Rothe, and Melanie Schienle (2012). "Nonparametric regression with nonparametrically generated covariates". In: *The Annals of Statistics* 40.2, pp. 1132–1170.

Masry, Elias (1996). "Multivariate local polynomial regression for time series: uniform strong consistency and rates". In: *Journal of Time Series Analysis* 17.6, pp. 571–599.

Miller, Amalia R (2011). “The effects of motherhood timing on career path”. In: *Journal of Population Economics* 24.3, pp. 1071–1100.

Neumeyer, Natalie and Holger Dette (2003). “Nonparametric comparison of regression curves: an empirical process approach”. In: *The Annals of Statistics* 31.3, pp. 880–920.

Reinhold, Steffen (2007). “Essays in demographic Economics”. PhD thesis. John Hopkins University.

Ribar, David C (1994). “Teenage fertility and high school completion”. In: *The Review of Economics and Statistics*, pp. 413–424.

Ruppert, David and Matthew P Wand (1994). “Multivariate locally weighted least squares regression”. In: *The Annals of Statistics*, pp. 1346–1370.

Sargan, John (1958). “The estimation of economic relationships using instrumental variables”. In: *Econometrica: Journal of the Econometric Society*, pp. 393–415.

van der Vaart, Aad and Jon Wellner (1996). *Weak Convergence and Empirical Processes*. Springer.

Vytlacil, Edward (2002). “Independence, monotonicity, and latent index models: An equivalence result”. In: *Econometrica* 70.1, pp. 331–341.

CHAPTER 2

Testing index sufficiency with a predicted index

2.1 Introduction

In many economic models, the outcome variable of interest depends on observed covariates only through a lower dimensional index. The notion that an index aggregates all information provided by the covariates is commonly referred to as *index sufficiency*. In this paper, I consider a statistical test of index sufficiency. In contrast to previous research on this topic, I do not assume that the rule that aggregates covariates into the index is known to the researcher. Instead, I assume that the rule is identified from the data and that a consistent estimator of the index is available.

My test uses a testing approach due to Delgado and Manteiga 2001 (henceforth cited as DM). The test statistic is based on the smoothed difference between the observed and the predicted outcome distribution. Replacing the true index by its estimated counterpart leads to a random perturbation of the smoother. In a regression setting, a comparable problem is known as the *generated regressors problem*. My main result fully accounts for the estimation error from estimating the index and gives an asymptotic expansion of the test statistic around its oracle, i.e., around a version of the test statistic in which the index is estimated without error.

Let X denote a random vector of observed covariates that takes values in $\mathbb{R}^{d_x}$. Let $r_0 : \mathbb{R}^{d_x} \to \mathbb{R}^d$ denote the true index and let Y denote a real outcome variable. Formally, I want to test the hypothesis

$$H_0 : E[Y \mid X] = E[Y \mid r_0(X)]. \tag{2.1}$$

So far, the literature has focused primarily on the cases $r_0(X) = X'\beta$ and $r_0(X) = v(X'\beta)$ for an unknown finite-dimensional parameter β and a possibly unknown nonparametric link function v. This is the classical single-index specification (Ichimura 1993). Model checks for this specification have been considered in Xia et al. 2004; Stute and Zhu 2005; Chen and Van Keilegom 2009; Escanciano and Song 2010 and Maistre and Patilea

2014. This paper adds to the literature by placing no restrictions on r_0 beyond general smoothness assumptions. In particular, r_0 can be a multi-index, i.e. it is allowed to map into vectors, and it can be fully nonparametric. I do not assume a specific first-stage estimator. Instead, I follow the recent literature on nonparametrically generated regressors (Mammen, Rothe, and Schienle 2012; Escanciano, Jacho-Chávez, and Lewbel 2014; Mammen, Rothe, and Schienle 2015) and provide high-level results that hold for a large class of estimators that satisfy certain accuracy and complexity assumptions. In terms of the scope of testable hypotheses, this paper is closest to Chen and Van Keilegom 2009 who propose an empirical likelihood test for the validity of a general class of semiparametric multi-index models. In terms of methodology, the present paper is closer to Xia et al. 2004 and Stute and Zhu 2005 who also use a testing approach based on empirical processes.

The hypothesis (2.1) is a special case of the significance testing problem considered in DM. Their testing procedure has very favorable asymptotic properties. In particular, it can detect local alternatives at a parametric rate, even though the null model is allowed to belong to a nonparametric class. Moreover, the asymptotic behavior of the test is independent of the size of X, allowing for a rich class of alternatives. However, for the problem considered in this paper, the test statistic employed by DM is unfeasible since the true index r_0 is unknown. I assume that an estimator $\hat{r}$ of r_0 is available and consider a feasible version of the test statistic in DM in which r_0 is replaced by $\hat{r}$.

This change is not innocuous. First, as I demonstrate below, the first-stage estimation of the index changes the asymptotic behavior of the test statistic. In particular, a rejection rule based on critical values computed according to the procedure suggested in DM will typically not control the type-I error of the test. Secondly, for the automatic bias removal that is built into the test statistic of DM to work well, the density of the estimated index has to lie in a similar smoothness class as the density of the true index. In various relevant scenarios such a condition does not hold. To account for this, I suggest an alternative procedure that estimates the bias and removes it from the test statistic. I show that this procedure works well under a set of conditions that do not restrict the density of the estimated index. As an added bonus, this procedure requires weaker assumptions about the order of the kernel function than the procedure in DM does. A third way in which using a predicted index affects the test is of a more theoretical nature. The arguments in DM rely on uniform convergence over a Vapnik-Chervonenkis (henceforth VC) class. This uniformity can be achieved under minimal assumptions about the smoothing parameter. All that is required is that the number of local observations approaches infinity. To account for a predicted index, my uniform convergence arguments have to deal with classes that are considerably more complex than VC classes. Consequently, my results impose more stringent assumptions on the smoothing parameter than DM do. However, for a wide range of testing problems, many bandwidth choices, including bandwidths that are optimal with respect to the mean integrated squared error (MISE), are still permissible.

The main building block of the stochastic expansion presented in this paper is a new lemma for kernel-weighted U-statistics (Lemma 2.5 in the Appendix). This lemma bounds the maximal deviation of the U-statistic using estimated kernel weights from the

U-statistic using the true kernel weights. I adopt accuracy and complexity assumptions about the predicted index that are similar to the assumptions imposed on generated regressors in Mammen, Rothe, and Schienle 2012. As expected, for U-statistics the conditions for ignoring higher-order effects of the predicted index are much weaker than the conditions required for ignoring higher-order effects for the empirical process considered in Mammen, Rothe, and Schienle 2012. I suspect that this result transcends my particular application and will prove its usefulness in other semiparametric estimation and testing problems with predicted quantities.

2.2 Motivating examples

In this section I present examples illustrating the relevance of index restrictions in empirical economic research.

Example 1. *Semiparametric binary choice model.* The model by Klein and Spady 1993 models the relationship between a covariate vector X taking values in $\mathbb{R}^{d_x}$ and a binary outcome Y. Their model features a index function $v(\cdot, \theta_0)$ which is known up to the finite dimensional parameter θ_0 and an infinite dimensional link function. The index function maps onto the real line and subsumes all information about the outcome, i.e.,

$$E[Y \mid X] = E[Y \mid v(X, \theta_0)].$$

Setting $r_0(x) = v(x, \theta_0)$ this equation is equivalent to hypothesis (2.1). In this example the index has dimension $d = 1$. Klein and Spady 1993 suggest a semiparametric likelihood estimator $\hat{\theta}$ that consistently estimates θ_0. Let $\hat{v}$ denote the function v with the true link function replaced by the nonparametric estimator suggested in Klein and Spady 1993. This yields $\hat{r}(x) = \hat{v}(x, \hat{\theta})$ as an obvious plug-in estimator of the unknown index.

Example 2. *Instrument validity in a treatment model.* Consider a setting with binary treatment D and latent outcomes Y_0 and Y_1. Let $Y = DY_1 + (1 - D)Y_0$ denote the observed outcome. Suppose that the econometrician observes a vector of instruments S taking values in $\mathbb{R}^{d_s}$ and a vector W of (other) covariates taking values in $\mathbb{R}^{d_w}$. Let $P(s, w) = E[D \mid S = s, W = w]$ denote the propensity score. Moreover, suppose that, as in Vytlacil 2002, $U \sim \text{Uniform}[0, 1]$ denotes an unobservable component that monotonically affects the selection into treatment and that

$$(Y_0, Y_1, U) \perp\!\!\!\perp S \mid W.$$

The latter condition imposes instrument validity in the sense that the variation of the instrument is required to be orthogonal to the variation of the unobservables. Following Dzemski and Sarnetzki 2014 it can be shown that

$$E[Y \mid S, W] = E[Y \mid P(S, W), W].$$

This restriction fits into the class of index sufficiency problems considered in this paper. Set $X = (S', W')'$ and let x denote a generic realization from X. Defining the index

$$r_0(x) = \big(P((x_1, \ldots, x_{d_s})', (x_{d_s+1}, \ldots, x_{d_s+d_w})'), x_{d_s+1}, \ldots, x_{d_s+d_w}\big)'$$

this equation can be rewritten into the format of equation (2.1). In this example the index has dimension $d = 1 + d_w$ and the unrestricted conditioning set has dimension $d_x = d_s + d_w$.

2.3 Test statistic

Suppose that the index $r_0(X)$ is continuously distributed and denote its density by $f_{r_0(X)}$. Moreover, define the population mean function $m(t) = \mathrm{E}[Y \mid r_0(X) = t]$. DM show that the hypothesis (2.1) is equivalent to

$$f_{r_0(X)}(r_0(X))E\big[Y - m(r_0(X)) \mid X\big] = 0 \quad \text{almost surely.}$$

With $\delta_x(\cdot) = \mathbf{1}_{\{\cdot \le x\}}$ this in turn can be rewritten in terms of unconditional moments as

$$T(x) = E\big[f_{r_0(X)}(r_0(X))(Y - m(r_0(X))\delta_x(X)\big] = 0 \quad \text{for } P_X\text{-almost all } x.$$

T is an unobserved population quantity and has to be estimated in order to be used in a test. To this end, suppose that a sample $(Y_i, X_i)_{1\le i\le n}$ from (Y, X) is available. There is also an estimator $\hat{r}$ of r_0 that may be correlated with the observed sample. Since estimating the index r_0 is a prerequisite for computing an estimator of T, $\hat{r}$ will be referred to as the first-step estimator. For a multivariate kernel function $K : \mathbb{R}^d \to \mathbb{R}$ and a bandwidth tupel $h = (h_1, \ldots, h_d)$ let

$$K_{h,ij}(r) = h_+^{-1} K\left(\frac{r(X_i) - r(X_j)}{h}\right)$$

where $h_+ = h_1 \cdots h_d$. Local constant estimators of $f_{r_0(X)}$ and m evaluated at $r_0(X_i)$ are given by

$$\hat{f}_{r_0(X),i} = \frac{1}{n-1}\sum_{j:j\ne i} K_{h,ij}(\hat{r}) \quad \text{and} \quad \hat{m}_i = \left(\hat{f}_{r_0(X),i}\right)^{-1}\frac{1}{n-1}\sum_{j:j\ne i} K_{h,ij}(\hat{r})Y_j.$$

Carrying over the estimator of T suggested in DM to a setup with a predicted index yields the estimator

$$\begin{aligned}\hat{T}_n(x) = T_n(\hat{r}, x) &= \frac{1}{n}\sum_i \hat{f}_{r_0(X),i}\big(Y_i - \hat{m}_i\big)\delta_x(X_i)\\ &= \frac{1}{n(n-1)}\sum_{i\ne j} K_{h,ij}(\hat{r})(Y_i - Y_j)\delta_x(X_i).\end{aligned}$$

A suitable test statistic measures the distance of $\hat{T}_n$ from the zero function and the corresponding decision rule rejects for large values of the test statistic. This approach exploits the fact that the population counterpart of $\hat{T}_n$ is identical to the zero function under the null hypothesis. Test statistics suggested by DM are

$$C_n = \int \left(\sqrt{n}\hat{T}_n\right)^2 dP_n \quad \text{and} \quad K_n = \sup_{x\in\mathbb{R}^{d_x}} \left|\sqrt{n}\hat{T}_n(x)\right|.$$

These are known as the Cramér-von Mises statistic and the Kolmogorov-Smirnov statistic, respectively.

2.4 Main results

The test does not place any parametric restrictions on the data generating process. Instead, the data generating process is assumed to belong to a smooth class. The first item of the following set of assumptions details this class. The second and third item give the assumed properties of the kernel smoother.

Assumption 2.1

i) (Smoothness) The mean function m has $(q+1)$ bounded derivatives with $q \geq 1$. The true index r_0 admits a bounded density $f_{r_0(X)}$ with q_2 bounded derivatives, $1 \leq q_2 \leq q$.

ii) (Kernel) Let K denote a product kernel, for $u \in \mathbb{R}^d$, $K(u) = \prod_{l=1}^{d} k(u_l)$ for a function $k : \mathbb{R} \to \mathbb{R}$. There is a constant $L > 0$ such that the function k satisfies for all $u_l, u_{l,1}, u_{l,2} \in \mathbb{R}$

$$\begin{aligned} |u_l| > L \Rightarrow k(u_l) = 0 \qquad & \textit{(bounded support)} \\ |k| \leq L \qquad & \textit{(boundedness)} \\ \left|k(u_{l,1}) - k(u_{l,2})\right| \leq L \left|u_{l,1} - u_{l,2}\right| \qquad & \textit{(Lipschitz continuity).} \end{aligned}$$

Moreover there is an integer $q_1 \geq 1$ such that

$$\begin{aligned} \int k(t)\, dt &= 1 \\ \int k(t) t^p \, dt &= 0 \quad \textit{for } p = 1, \ldots, q_1. \end{aligned}$$

iii) (Bandwidth) Write the bandwidth tupel as $h = (h_1, \ldots, h_d)$. For each $j = 1, \ldots, d$, there is a $\eta_j > 0$ such that $h_j \asymp n^{-\eta_j}$. Let $\eta_+ = \eta_1 + \cdots + \eta_d$ and $h_+ = h_1 \cdots h_d$ so that $h_+ \asymp n^{-\eta_+}$.

Item (i) posits that the regression function m behaves locally like a polynomial. In conjunction with a similar assumption on the density function $f_{r_0(X)}$ this implies that kernel-weighed expectations of the regression function can be well approximated by a polynomial in the moments of the kernel function. In conjunction with the assumption of a higher-order kernel in item (ii) this allows for a smoother that takes out lower-order bias terms. Finally, item (iii) mandates that the bandwidths vanish at a polynomial rate.

The results in this paper do not presume a particular first-stage estimator. Following Mammen, Rothe, and Schienle 2012 I assume instead that $\hat{r}$ belongs to a large class of uniformly consistent estimators. The restrictions placed on this class are summarized by the following assumption.

Assumption 2.2 (Estimator of index function)
For each $j = 1, \ldots, d$ there is a $\delta_j > \eta_j$ and a sequence of sets $\tilde{\mathcal{R}}_{n,j}$ such that with probability approaching one $\hat{r}_j$ is contained in the class

$$\mathcal{R}_{n,j} = \left\{ r_j \in \tilde{\mathcal{R}}_{n,j} : \|r_j - r_{0,j}\|_\infty \leq n^{-\delta_j} \right\},$$

i.e., $\lim_{n\to\infty} P(\hat{r}_j \in \mathcal{R}_{n,j}) = 1$. *For each* $j = 1,\dots,d$, *there are a* $\gamma_j \in [0,1)$ *and a* $\xi_j > 0$ *such that* $\mathcal{R}_{n,j}$ *can be covered by (less than)* $\exp(n^{\xi_j} u^{-\gamma_j})$ *u-balls with respect to the* $\|\cdot\|_\infty$*-metric. Let* $\mathcal{R}_n = \mathcal{R}_{n,1} \times \cdots \times \mathcal{R}_{n,d}$.

This assumption imposes two kinds of restrictions on the class that contains $\hat{r}$. First, it assumes a uniform rate of consistency. This is a statement about the *precision* of the estimator. The condition $\delta_j > \eta_j$ ensures that the predicted index is locally informative. When taking local averages it guarantees that, based on the predicted index, the smoother can correctly identify observations that are close to each other in terms of the true index. Secondly, the assumption restricts the *complexity* of the class by imposing an upper bound on its covering or entropy number. Under this restriction I can derive uniform expansions, i.e., expansions that are true for all realizations of the predicted index. This approach allows me to procede without imposing any restrictions on the correlation structure between the predicted index and the observed sample. As discussed in Mammen, Rothe, and Schienle 2012 the complexity assumption can be verified for a diverse range of estimators such as parametric estimators, series estimators or kernel-based estimators. In Section 2.5, I provide a detailed discussion of how to apply my results when a local polynomial estimator is used to estimate the index.

To analyze the behavior of $\hat{T}_n(x)$ it is convenient to split off the bias term. Define the observational error $\epsilon = Y - m(r_0(X))$. Under the null hypothesis of index sufficiency $E[\epsilon \mid X] = 0$. To separate error and bias term write

$$\begin{aligned}\hat{T}_n(x) =& \frac{1}{n(n-1)} \sum_{i \neq j} K_{h,ij}(\hat{r})(\epsilon_i - \epsilon_j)\delta_x(X_i) \\ &+ \frac{1}{n(n-1)} \sum_{i \neq j} K_{h,ij}(\hat{r})\big[m(r_0(X_i)) - m(r_0(X_j))\big]\delta_x(X_i) \\ =& \hat{T}_{\text{error},n}(x) + \hat{T}_{\text{bias},n}(x).\end{aligned}$$

Under relatively weak conditions, the estimation of the index affects the asymptotic behavior of $\hat{T}$ only through the bias term. For each $x \in \mathbb{R}^{d_x}$ let $\mu_x : \mathbb{R}^d \to \mathbb{R}$ be given by $s \mapsto P[X \leq x \mid r_0(X) = s]$. Collect these functions in the class $\mathcal{M} = \{\mu_x : x \in \mathbb{R}^d\}$. DM give conditions under which

$$\hat{T}_{\text{error},n}(x) \approx \frac{1}{n} \sum_i f_{r_0(X)}(r_0(X_i))\big(\delta_x(X_i) - \mu_x(r_0(X_i))\big)\epsilon_i.$$

It turns out that a similar result is true if r_0 is replaced by its estimator $\hat{r}$.

Theorem 2.1 *Suppose that there is a* $s > 2$ *such that* $E|\epsilon|^s < \infty$ *and*

$$\begin{aligned}&0 < \frac{1}{2}\big[1 - \eta_+\big] + (\delta - \eta)_{\min} - \max_{1\leq \ell \leq d}\big[\delta_\ell \gamma_\ell + \xi_\ell\big] \\ &0 < (\delta - \eta)_{\min} - \frac{1}{2}\max_{1\leq \ell \leq d}\big[\delta_\ell \gamma_\ell + \xi_\ell\big] \\ &0 < 1 - \eta_+\end{aligned}$$

and assume that the class $\mathcal{M}$ satisfies a uniform Lipschitz condition. Under Assumption 2.1 and Assumption 2.2 we have

$$\sup_{x\in\mathbb{R}^{d_x}} \left|\hat{T}_{error,n}(x) - \frac{1}{n}\sum_i f_{r_0(X)}(r_0(X_i))\big(\delta_x(X_i) - \mu_x(r_0(X_i))\big)\epsilon_i\right| = o_p\big(n^{-\frac{1}{2}}\big).$$

In addition to conditions similar to those imposed by DM, this result requires some restrictions on the behavior of the estimator $\hat{r}$. The stringency of these restrictions depends on the precision of the least precisely estimated component of $\hat{r}$ and on the complexity of the most complex component of $\hat{r}$. Notably, the restrictions do not depend on the size of the index vector. The restrictions on the first-stage estimator are intertwined with the choice of the bandwidth h. Choosing larger bandwidths (i.e. smaller η_+) admits a larger class of estimators. This is quite intuitive. For a given first-stage estimator, choosing larger bandwidths reduces the complexity of the estimated kernel weights and thus makes it easier to derive an uniform expansion.

Theorem 2.1 considers a U-statistic of order two and gives conditions under which asymptotic effects of estimating the index are negligible. It is instructive to compare these with the conditions required in Lemma 1 of Mammen, Rothe, and Schienle 2012 which considers the corresponding problem for an empirical process, i.e., a U-statistic of order one. They give the condition

$$\frac{1}{2} < \frac{1}{2}(1-\eta_+) + (\delta-\eta)_{\min} - \frac{1}{2}\max_{1\le\ell\le d}\big[\delta_\ell\gamma_\ell + \xi_\ell\big].$$

With much to spare, this is always a more stringent assumption than the second condition in Theorem 2.1. It is more restrictive than the first condition in Theorem 2.1 for classes that satisfy $\max_{1\le\ell\le d}\left[\delta_\ell\gamma_\ell + \xi_\ell\right] < 1$.

Next, consider the bias term $\hat{T}_{\text{bias},n}(x)$. Using multi-index notation (see Appendix 2.A) we can Taylor-expand the regression function m and write

$$m(r_0(X_i)) - m(r_0(X_j)) \approx -\sum_{1\le|\alpha|\le q}\frac{1}{\alpha!}\partial^\alpha m(r_0(X_i))\big(r_0(X_j) - r_0(X_i)\big)^\alpha.$$

In order to use this expression in the characterization of the bias term it is convenient to define $P_n(r,x) = \sum_{1\le|\alpha|\le q} P_n^\alpha(r,x)$ with

$$P_n^\alpha(r,x) = \frac{1}{n(n-1)}\sum_{i\neq j}\delta_x(X_i)\,K_{h,ij}(r)\,\frac{h^\alpha}{\alpha!}\,\partial^\alpha m(r_0(X_i))\left(\frac{r(X_j)-r(X_i)}{h}\right)^\alpha.$$

In the setting of DM the bias term $\hat{T}_{\text{bias},n}(x) \approx 0$. The following theorem indicates that this is no longer true if the index is estimated.

Theorem 2.2 *Suppose that Assumption 2.1 and Assumption 2.2 hold. For $q^* \le q$ let*

$$\Delta(r,x) = \frac{1}{n(n-1)}\sum_{i\neq j}E\Bigg[\delta_x(X_i)K_{h,ij}(r_0)\sum_{1\le|\alpha|\le q^*}\frac{1}{\alpha!}\partial^\alpha m(r_0(X_i))\big\{\,(r(X_j)-r(X_i))^\alpha - (r_0(X_j)-r_0(X_i))^\alpha\,\big\}\Bigg]$$

and $\Xi(r,x) = EP_n(r,x)$. *Suppose that*

$$\begin{aligned}
&0 < \frac{1}{2}\big(1-\eta_+\big) + (\delta-\eta)_{\min} + \eta_{\min} - \max_{1\le\ell\le d}\left[\delta_\ell\gamma_\ell + \xi_\ell\right]\\
&0 < (\delta-\eta)_{\min} + \eta_{\min} - \frac{1}{2}\max_{1\le\ell\le d}\left[\delta_\ell\gamma_\ell + \xi_\ell\right]\\
&\frac{1}{2} < (\delta-\eta)_{\min} + (q^*+1)\eta_{\min}\\
&\frac{1}{2} < (\delta-\eta)_{\min} + \delta_{\min}\\
&\frac{1}{2} < (q+1)\eta_{\min}.
\end{aligned}$$

Then

$$\sup_{x\in\mathbb{R}^{d_x}} \left|\hat{T}_{bias,\ n}(x) - \Delta(\hat{r},x) + \Xi(\hat{r},x)\right| = o_p\big(n^{-\frac{1}{2}}\big).$$

If one is allowed to choose $q^* = 1$ and if a linear representation of the first-stage estimator is available then it is fairly straightforward to characterize the behavior of $\Delta(\hat{r},x)$. This is illustrated by the example discussed in Section 2.5. Choosing $q^* > 1$ relaxes the restrictions on the first-stage estimator at the expense of a more involved structure of $\Delta(\hat{r},x)$.

The term $\Xi(\hat{r},x)$ in Theorem 2.2 is the usual bias term in nonparametric kernel regression. DM adopt smoothness assumptions under which its lower-order terms are removed by using the higher-order kernel smoother. Their assumption can be adopted to the setting with a predicted index.

Assumption 2.3
Suppose that Assumption 2.1 holds with $q_1 = q_2 = q$. *Moreover, suppose that each* $r \in \mathcal{R}_n$ *admits a density* $f_{r(X)}$ *and suppose that all densities in* $\{f_{r(X)} : r \in \mathcal{R}_n\}$ *have* q *derivatives that are bounded uniformly over this class.*

Under this assumption $\Xi(\hat{r},x) = O_p\big(n^{-(q+1)\eta_{\min}}\big)$ uniformly in $x \in \mathbb{R}^{d_x}$. In addition to restricting the population density as DM do this assumption also imposes restrictions on the estimator. The density of an estimator does not typically inherit the smoothness properties of its population counterpart. In some cases Assumption 2.3 will be violated even though the population density $f_{r_0(X)}$ may be perfectly smooth. For settings in which the validity of Assumption 2.3 is in doubt I suggest an alternative procedure. Suppose that an estimator $\hat{\Xi}(x)$ of $\Xi(\hat{r},x)$ is available and define the bias-corrected term

$$\hat{T}^{\text{corr}}_{\text{bias, n}}(x) = T_{\text{bias, n}}(x) - \hat{\Xi}(x).$$

The challenge in constructing such an estimator is that it has to converge at a rate that is faster than the parametric rate to ensure that

$$\sup_{x\in\mathbb{R}^{d_x}} \left|\hat{T}^{\text{corr}}_{\text{bias, n}}(x) - \Delta(\hat{r},x)\right| = o_p\big(n^{-\frac{1}{2}}\big).$$

As it turns out, the conditions for such an estimator to exist are not too strong. The proof of the following result exploits the fact that a higher-order kernel reduces bias under an oracle version of the test.

Theorem 2.3 *Suppose that Assumption 2.1 and Assumption 2.2 hold and suppose that* $r_0(X)$ *has compact support. There are positive constants* $\theta_1, \dots, \theta_q$ *and estimators* $\widehat{\partial^\alpha m}$ *such that*

$$\sup_t \left|\widehat{\partial^\alpha m}(t) - \partial^\alpha m(t)\right| = O_p\Big(n^{-\theta_{|\alpha|}}\Big)$$

for all $1 \leq |\alpha| \leq q$. *Define the estimator* $\hat{\Xi} = \sum_{1\leq|\alpha|\leq q} \hat{\Xi}^\alpha$ *with*

$$\hat{\Xi}^\alpha(x) = \frac{1}{n(n-1)} \sum_{i\neq j} \delta_x(X_i)\, K_{h,ij}(\hat{r})\, \frac{h^\alpha}{\alpha!}\, \widehat{\partial^\alpha m}(\hat{r}(X_i)) \left(\frac{\hat{r}(X_j) - \hat{r}(X_i)}{h}\right)^\alpha .$$

Let $\kappa = \min\{\kappa_1, \dots, \kappa_7\}$ *for*

$$\begin{aligned}
\kappa_1 &< \frac{1}{2}\big(1 - \eta_+\big) + \min_{1\leq s\leq q} \big\{\theta_s + s\eta_{\min}\big\} \\
\kappa_2 &< \frac{1}{2}\big(1 - \eta_+\big) + \delta_{\min} + \eta_{\min} \\
\kappa_3 &< (q_1 \wedge q_2 + 1)\eta_{\min} + \min_{1\leq s\leq q} \big\{\theta_s + s\eta_{\min}\big\} \\
\kappa_4 &< (q_1 \wedge q_2 + 1)\eta_{\min} + \delta_{\min} + \eta_{\min} \\
\kappa_5 &< \frac{1}{2} + \frac{1}{2}\big(1 - \eta_+\big) + (\delta - \eta)_{\min} + \eta_{\min} - \max_{1\leq \ell\leq d} \left[\delta_\ell\gamma_\ell + \xi_\ell\right] \\
\kappa_6 &< \frac{1}{2} + (\delta - \eta)_{\min} + \eta_{\min} - \frac{1}{2} \max_{1\leq \ell\leq d} \left[\delta_\ell\gamma_\ell + \xi_\ell\right] . \\
\kappa_7 &< \frac{1}{2} + \eta_{\min}
\end{aligned}$$

Then

$$\sup_{x\in\mathbb{R}^{d_x}} \left|\hat{\Xi}(x) - \Xi(\hat{r}, x)\right| = o_p\big(n^{-\kappa}\big).$$

The results presented in this section imply an expansion of $\hat{T}_n(x)$ around its oracle $T_n(r_0, x)$. DM show that

$$T_n(r_0, x) \approx \frac{1}{n} \sum_i f_{r_0(X)}(r_0(X_i))\big(\delta_x(X_i) - \mu_x(r_0(X_i))\big)\epsilon_i.$$

Therefore, Theorem 2.1 and Theorem 2.2 give conditions under which

$$\hat{T}_n(x) - T_n(r_0, x) \approx \Delta(\hat{r}, x) - \Xi(\hat{r}, x).$$

If Assumption 2.3 holds or if $\hat{T}_n$ is suitably corrected for bias then $\Xi(\hat{r}, x)$ drops out of the expression in the previous display. Then, $\Delta(\hat{r}, x)$ can be interpreted as giving the deviation of $\hat{T}_n$ from its oracle. It summarizes the effect that the first-stage estimation has on the asymptotic distribution of the test statistic.

2.5 Application

In this section I consider a concrete example of a first-stage estimator and I apply the results from the previous section in order to quantify the influence of the first-stage estimation on the asymptotic distribution. Suppose that the index r_0 is identified from the moment equation

$$E[\breve{Y} \mid X] = r_0(X),$$

where $\breve{Y}$ is an observed outcome variable. Divide X into two subvectors $X^{(1)} \in \mathbb{R}^{d_x^{(1)}}$ and $X^{(2)} \in \mathbb{R}^{d_x^{(2)}}$, i.e., $d_x = d_x^{(1)} + d_x^{(2)}$ and $X = (X^{(1)}, X^{(2)})$. Suppose that r_0 depends only on $X^{(1)}$. In a slight abuse of notation, write $r_0(X) = r_0(X^{(1)})$. In this section, I consider local polynomial estimation of r_0. For expositional reasons, I discuss the simplest case possible, where both the covariate $X^{(1)}$ and the index $r_0(X^{(1)})$ have dimension one, i.e., $d = d_x^{(1)} = 1$. At the expense of a more involved notation, it is straightforward to adapt the results to cover more convoluted cases. In Appendix 2.D, I discuss the validity of Assumption 2.2 for general local polynomial estimators without any restrictions on the dimensionality. In the following, I assume that the first-stage estimator $\hat{r}$ and the second-stage estimator $\hat{T}_n$ are computed from the same sample. In particular, the procedure considered here does not require the researcher to split the sample.

Formally, suppose that a sample $(Y_i, \breve{Y}_i, X_i^{(1)}, X_i^{(2)})_{1 \le i \le n}$ from $(Y, \breve{Y}, X^{(1)}, X^{(2)})$ is available and let $\hat{T}_n$ denote the estimator defined in Section 2.3 computed on this sample. Let $L : \mathbb{R}^{d_x^{(1)}} \to \mathbb{R}$ denote a uni-variate kernel function. Also, let g denote a bandwidth sequence and suppose that $g \asymp n^{-\eta^*}$. Write $L_g(\cdot) = g^{-1} L(\cdot/g)$. The p-th order local polynomial estimator of $r_0(x)$ is given by $\hat{r}(x) = \hat{b}_0(x)$ where

$$(\hat{b}_j(x))_{0 \le j \le p} \in \arg\min_{b_j, 0 \le j \le p} \sum_i \left[\breve{Y}_i - \sum_{0 \le j \le p} b_j \left(X_i^{(1)} - x^{(1)} \right)^j \right]^2 L_g \left(X_i^{(1)} - x^{(1)} \right)$$

The following assumption summarizes some regularity assumptions about the distribution of X.

Assumption 2.4
Suppose that the random variable $X^{(1)}$ takes values in a compact set $\mathcal{D} \subset \mathbb{R}^{d_{x^{(1)}}}$ and that its density $f_{X^{(1)}}$ is bounded away from zero on $\mathcal{D}$.

This assumption is fairly strong and chosen mainly to allow for an accessible presentation. At the expense of additional notational clutter they can be relaxed considerably. In particular, the assumption that the density $f_{X^{(1)}}$ is bounded away from zero is non-essential. A key feature of the expansion from Theorem 2.2 is that it depends on a smoothed version of the first-stage estimator. The region where the density $f_{X^{(1)}}$ is small and where the local polynomial estimator behaves irregularly is automatically weighed down. The down-weighing is by a locally smoothed version of the density rather than the density itself and can therefore not offset irregular behavior completely. However, the set-up allows for densities that approach zero as the sample size increases. As is apparent from the proof of Theorem 2.4 the density is allowed to approach zero at a rate $n^{-\tau}$, $0 < \tau < \eta^*$. This can be exploited by introducing an explicit trimming sequence that trims out observations in the region where the density takes small values.

The following result describes the behavior of $\hat{T}_n$ when the first-stage estimator is chosen to be a local linear estimator, i.e., when the order of the local polynomial is $p = 1$. I assume that the bandwidth sequences g and h vanish at MISE-optimal rates. It is straightforward to extend this result to allow for a higher-order polynomial, bandwidths that are not MISE-optimal or even to allow for higher-dimensional indices ($d > 1$) or indices that aggregate higher dimensional covariates ($d_{x^{(1)}} > 1$). An advantage of using MSE-optimal rates is that these are the rates targeted by data driven bandwidth selection procedures (Jones, Marron, and Sheather 1996; Härdle and Marron 1985). Therefore, it is possible to implement a test that satisfies the assumptions of the theorem.

Theorem 2.4 *Suppose that Assumption 2.1 holds with $d = d_x = 1$ and $q = 2$, and assume that the class $\mathcal{M}$ satisfies a uniform Lipschitz condition. Moreover, Assumption 2.4 holds and the first-stage kernel L has bounded support and one Lipschitz continuous derivative. The bandwidth sequences h and g are chosen to be MISE-optimal. Let $\zeta = E[\check{Y} - r_0(X) \mid X]$ and suppose that $E|\zeta|^2 < \infty$ and $E|\epsilon|^s < \infty$ for a $s > 0$. If the first stage estimator fits a local polynomial of order $p = 1$ (local linear estimator) then,*

$$\sup_{x \in \mathbb{R}^{d_x}} \left| \hat{T}_n(x) - \Xi(\hat{r}, x) - \left\{ \tilde{T}_n(x) + \tilde{B}_n(x) \right\} \right| = o_p\left(n^{-\frac{1}{2}}\right),$$

where

$$\begin{aligned} \tilde{T}_n(x) =& \frac{1}{n} \sum_i f_{r_0(X)}(r_0(X_i)) \left\{ \delta_x(X) - \mu_x(r_0(X)) \right\} \left[\epsilon_i - \partial m(r_0(X_i)) \zeta_i \right], \\ \tilde{B}_n(x) =& - E\left[f_{r_0(X)}(r_0(X)) \left\{ \delta_x(X) - \mu_x(r_0(X)) \right\} \partial m(r_0(X)) B_n^*(X) \right], \end{aligned}$$

and B_n^* is a bias term defined in Appendix 2.D.

$\tilde{T}_n(x)$ given above is reminiscent of the quantity $\tilde{U}_n$ defined in DM. Following their arguments, it can be shown that $\tilde{T}_n$ converges weakly to a zero-mean Gaussian process indexed by $x \in \mathbb{R}^{d_x}$ with covariance kernel

$$\begin{aligned} \Sigma_{x,x'} = E\Big[& f^2_{r_0(X)}(r_0(X)) [\delta_x(X) - \mu_x(r_0(X))] [\delta_{x'}(X) - \mu_{x'}(r_0(X))] \\ & \left\{ \sigma^2_\epsilon(X) - 2 \partial m(r_0(X)) \sigma_{\epsilon,\zeta}(X) + [\partial m(r_0(X))]^2 \sigma^2_\zeta(X) \right\} \Big], \end{aligned}$$

where $\sigma^2_\epsilon(x) = E[\epsilon^2 \mid X = x]$, $\sigma^2_\zeta(x) = E[\zeta^2 \mid X = x]$ and $\sigma_{\epsilon,\zeta}(x) = E[\epsilon\zeta \mid X = x]$.

Since $\tilde{T}_n$ is unobserved, it can not serve as a basis to conduct inference. However, it is possible to construct a muliplier-type bootstrap version of $\tilde{T}_n$. Sample independently of $(Y_i, \check{Y}_i, X_i)_{1 \le i \le n}$ from a random variable V with $EV = 0$ and $EV^2 = 1$. Denote this sample by $(V_i)_{1 \le i \le n}$. The bootstrap version of $\tilde{T}_n$ is given by

$$\tilde{T}_n^*(x) = \frac{1}{n} \sum_i \widehat{f_{r_0(X)}}(\hat{r}(X_i)) \left(\delta_x(X_i) - \widehat{\mu_x}(\hat{r}(X_i)) \right) \left[\hat{\epsilon}_i - \partial m \widehat{(r_0(X_i)})) \hat{\zeta}_i \right] V_i,$$

where $\widehat{f_{r_0(X)}}$, $\widehat{\mu_x}$ and $\partial m \widehat{(r_0(X_i)}))$ are appropriate estimators and $\hat{\epsilon}_i$ and $\hat{\zeta}_i$ are the empirical residuals.

2.6 Conclusion

Section 2.5 illustrates how the high-level results in Section 2.4 can be used to construct a test for a concrete testing problem. Similarly, one can construct tests for other models that impose index sufficiency. Some motivating examples are presented above. In addition, Chen and Van Keilegom 2009 and Maistre and Patilea 2014 discuss several such models that figure prominently in the econometric literature.

For ease of exposition, I have focused on real-valued outcomes. The apporach described in this paper can be extended to also cover outcome variables that are vectors.

This paper focuses on index-sufficiency in the mean. As discussed in Maistre and Patilea 2014, the researcher might be interested in testing the stronger statement of index-sufficiency in the conditional law, i.e.,

$$H_0' : Y \perp\!\!\!\perp X \mid r_0(X).$$

This is equivalent to a problem of testing infinitely many mean restrictions

$$H_0'' : E[\mathbf{1}_{\{Y \le y\}} \mid X] = E[\mathbf{1}_{\{Y \le y\}} \mid r_0(X)] \qquad \text{for all } y \in \mathbb{R}.$$

Upon cursory inspection it seems that it is relatively straightforward to extend the results from Section 2.4 to hold uniformly over a class of outcomes indexed by a VC-class. Therefore, it may be possible to employ an extension of my approach to test for index-sufficiency in the conditional law. To my knowledge, properties of such a test are currently unknown and further research is needed.

Appendix 2.A Notation

For $p > 0$ and a measure μ let $\|\cdot\|_{\mu,p}$ denote the L_p norm with respect to μ, i.e., for all μ-measurable functions f we take $\|f\|_{\mu,p} = \int |f|^p \, d\mu$. Take $\|\cdot\|_\mu = \|\cdot\|_{\mu,2}$. If argument θ and the domain Θ of a function f are obvious, the supremum norm $\sup_{\theta\in\Theta} |f(\theta)|$ is denoted by $\|f\|_\infty$. P always denotes the population probability measure and E denotes the corresponding expectation. For a random variable Z, P_Z denotes the measure that "integrates Z out", i.e., that integrates with respect to the marginal distribution of Z. P_n denotes the empirical measure. U_n denotes the empirical U-statistic of order 2, i.e., the measure that assigns mass $n(n-1)$ to each ordered pair of observations. The covering or entropy number of a class $\mathcal{G}$ for a cover consisting of u-balls in the $\|\cdot\|$-norm is denoted $\mathcal{N}(u, \mathcal{G}, \|\cdot\|)$. To write out multivariate Taylor expansions, I use multi-index notation. For a multi-index $\alpha \in \mathbb{N}_0^d$, $x \in \mathbb{R}^d$ and appropriately differentiable $f : \mathbb{R}^d \to \mathbb{R}$ let

$$\begin{aligned} \alpha! &= \alpha_1! \cdots \alpha_d! & \qquad |\alpha| &= \alpha_1 + \cdots + \alpha_d \\ \partial^\alpha f &= \partial_1^{\alpha_1} f \cdots \partial_d^{\alpha_d} f & \qquad x^\alpha &= x_1^{\alpha_1} \cdots x_d^{\alpha_d}. \end{aligned}$$

For vectors $v, v_1, v_2 \in \mathbb{R}^d$ let $\mathrm{Diag}(v)$ denote the $d \times d$ diagonal matrix with the vector v on the diagonal and let ${}^{v_1}\!/_{v_2} = [\mathrm{Diag}(v_2)]^{-1} v_1$. Moreover, let $v_{\min} = \min_{1\le i\le d} v_i$. For a bandwidth vector $h \in \mathbb{R}^d_{++}$ let $h_+ = \prod_{j=1}^d h_j$. For a kernel K, bandwidth vector h and observations i and j kernel weights are defined by

$$K_{h,ij}(r) = h_+^{-1} K\left(\frac{r(X_i) - r(X_j)}{h}\right).$$

Appendix 2.B Proofs of main theorems

Proof of Theorem 2.1 The proof starts out from the following decomposition

$$\begin{aligned}
&\sup_{x\in\mathbb{R}^{d_x}} \Big|\frac{1}{n(n-1)} \sum_{i\neq j} \delta_x(X_i) K_{h,ij}(\hat{r})\big(\epsilon_i - \epsilon_j\big) \\
&\qquad - \frac{1}{n}\sum_i f_{r_0(X)}(r_0(X_i))\big(\delta_x(X_i) - \mu_x(r_0(X_i))\big)\epsilon_i\Big| \\
=& \sup_{x\in\mathbb{R}^{d_x}} \Big|\frac{1}{n(n-1)} \sum_{i\neq j} \delta_x(X_i)\big(K_{h,ij}(\hat{r}) - K_{h,ij}(r_0)\big)\epsilon_i\Big| \\
&+ \sup_{x\in\mathbb{R}^{d_x}} \Big|\frac{1}{n(n-1)} \sum_{i\neq j} \delta_x(X_i)\big(K_{h,ij}(\hat{r}) - K_{h,ij}(r_0)\big)\epsilon_j\Big| \\
&+ \sup_{x\in\mathbb{R}^{d_x}} \Big|\frac{1}{n(n-1)} \sum_{i\neq j} \delta_x(X_i) K_{h,ij}(r_0)\epsilon_i - \frac{1}{n}\sum_i f_{r_0(X)}(r_0(X_i))\delta_x(X_i)\epsilon_i\Big| \\
&+ \sup_{x\in\mathbb{R}^{d_x}} \Big|\frac{1}{n(n-1)} \sum_{i\neq j} \delta_x(X_i) K_{h,ij}(r_0)\epsilon_j - \frac{1}{n}\sum_i f_{r_0(X)}(r_0(X_i))\mu_x(r_0(X_i))\epsilon_i\Big| \\
=& A_1 + A_2 + A_3 + A_4.
\end{aligned}$$

The goal is to prove that $A_1, \dots, A_4$ are all contained in the $o_p\big(n^{-\frac{1}{2}}\big)$-class. By Lemma 2.5 and Lemma 2.2

$$A_1 = \frac{1}{n}\sum_i E_{X_j}\big[K_{h,ij}(\hat{r}) - K_{h,ij}(r_0)\big]\delta_x(X_i)\epsilon_i + o_p\big(n^{-\frac{1}{2}}\big) = o_p\big(n^{-\frac{1}{2}}\big).$$

Similarly, one can show that $A_2 = o_p\big(n^{-\frac{1}{2}}\big)$. The arguments for A_3 and A_4 are similar. Here, I present only the proof for A_4 as it requires slightly more convoluted arguments than the corresponding proof for A_3. By Lemma 2.6

$$A_4 = \frac{1}{n}\sum_i \Big(E_{X_j}\big[K_{h,ij}(r_0)\delta_x(X_j)\big] - f_{r_0(X)}(r_0(X_i))\delta_x(X_i)\Big)\epsilon_i + o_p\big(n^{-\frac{1}{2}}\big).$$

Let

$$\begin{aligned}
G_{a,i}(x) &= E_X\left[h_+^{-1} K\left(\frac{r_0(X) - r_0(X_i)}{h}\right)\delta_x(X)\right]\epsilon_i \\
G_{b,i}(x) &= f_{r_0(X)}(r_0(X_i))\mu_x(r_0(X_i))\epsilon_i.
\end{aligned}$$

Show that

$$\sup_{x\in\mathbb{R}^{d_x}}\left|\frac{1}{n}\sum_i\big(G_{a,i}(x)-G_{b,i}(x)\big)\right|=o_p\big(n^{-\frac{1}{2}}\big).$$

Let $\mathcal{G}=\{G_{a,i}(x)-G_{b,i}(x):x\in\mathbb{R}^{d_x}\}$ and let $A_n=\{\|\epsilon_i\|_{P_n,s}\le C_A\}$ with C_A large enough that $PA_n\to 1$. There is a constant $\tilde{C}$ such that on A_n

$$\log\mathcal{N}\left(u,\mathcal{G},\|\cdot\|_{P_n}\right)\le\tilde{C}H_n(u)$$

for

$$H_n(u)=(\log n)+\log(u^{-1}\vee 1)$$

and all $u>0$. Work conditional on A_n. Construct a u-cover of $\mathcal{G}$ in the following way. Take $\tilde{u}=(C^*)^{-1}h_+(u/2)$. The class $\{\delta_x:x\in\mathbb{R}^{d_x}\}$ is VC (Pollard 1984, p. 18) and so is $\mu_x(r_0(X_i))=E[\delta_x(X_i)\mid r_0(X_i)]$ by Lemma 2.6.18 (vii) in van der Vaart and Wellner 1996. Let $Q=\frac{2s}{s-2}$. Take $\mathcal{C}_{1,n}\subset\mathbb{R}^{d_x}$ to be a minimal set such that $\{\delta_x:x\in\mathcal{C}_{1,n}\}$ covers $\{\delta_x:x\in\mathbb{R}^{d_x}\}$ with respect to the $\|\cdot\|_{P,Q}$-norm. Similarly, take $\mathcal{C}_{2,n}\subset\mathbb{R}^{d_x}$ to be a minimal set such that $\{\mu_x:x\in\mathcal{C}_{1,n}\}$ covers $\{\mu_x:x\in\mathbb{R}^{d_x}\}$ with respect to the $\|\cdot\|_{P_n,Q}$-norm. By the VC property of the two classes there exist constants A and V such that $\mathcal{C}_{1,n}$ and $\mathcal{C}_{2,n}$ can be chosen to contain less than $A\tilde{u}^{-V}$ elements each (Theorem 2.6.4 in van der Vaart and Wellner 1996). By construction

$$\mathcal{C}_n=\{G_{a,i}(x)-G_{a,i}(x'):x\in\mathcal{C}_{1,n},x'\in\mathcal{C}_{2,n}\}.$$

is a cover of $\mathcal{G}$ and can be chosen such that $\log\#\mathcal{C}_n\le\tilde{C}\,H_n(u)$ for a constant $\tilde{C}$ that is independent of u. Next, I show that if C^* is chosen large enough, then $\mathcal{C}_n$ will be a u-cover of $\mathcal{G}$ with respect to the $\|\cdot\|_{P_n}$-norm. Fix any $x_1\in\mathbb{R}^{d_x}$ and take $(x_2,x_2')\in\mathcal{C}_{1,n}\times\mathcal{C}_{2,n}$ to be a nearest grid-point, i.e., $\|\delta_{x_1}-\delta_{x_2}\|_{P,Q}\le\tilde{u}$ and $\left\|\mu_{x_1}-\mu_{x_2'}\right\|_{P_n,Q}\le\tilde{u}$. By Jensen's inequality

$$\begin{aligned}&\left|E_X\left[h_+^{-1}K\left(\frac{r_0(X)-r_0(X_i)}{h}\right)\big(\delta_{x_1}(X)-\delta_{x_2}(X)\big)\right]\right|^Q\\ \le&C\,h_+^{-Q}E_X\left|\delta_{x_1}(X)-\delta_{x_2}(X)\right|^Q.\end{aligned}$$

Therefore, by Hölder's inequality

$$\begin{aligned}\|G_{a,i}(x_1)-G_{a,i}(x_2)\|_{P_n}\le&C\,h_+^{-1}\left(E_X\left|\delta_{x_1}(X)-\delta_{x_2}(X)\right|^Q\right)^{\frac{1}{Q}}\left(\frac{1}{n}\sum_i|\epsilon_i|^s\right)^{\frac{1}{s}}.\\ \le&C^*\,h_+^{-1}\,\|\delta_{x_1}-\delta_{x_2}\|_{P,Q}\le\frac{u}{2}.\end{aligned}$$

Moreover,

$$\begin{aligned}&\left\|G_{b,i}(x_1)-G_{b,i}(x_2')\right\|_{P_n}\\ \le&C\,h_+^{-1}\left(\frac{1}{n}\sum_i\left|f_{r_0(X)}(r_0(X_i))\big(\mu_{x_1}(r_0(X_i))-\mu_{x_2'}(r_0(X_i))\big)\right|^Q\right)^{\frac{1}{Q}}\left(\frac{1}{n}\sum_i|\epsilon_i|^s\right)^{\frac{1}{s}}\\ \le&C^*\,h_+^{-1}\left\|\mu_{x_1}-\mu_{x_2'}\right\|_{P_n,Q}\le\frac{u}{2}.\end{aligned}$$

Collecting the results from above and applying the triangle inequality gives

$$\left\|G_{a,i}(x_1) - G_{b,i}(x_1) - (G_{a,i}(x_2) - G_{b,i}(x_2'))\right\|_{P_n} \le u$$

and thus confirms that $\mathcal{C}_n$ is indeed a u-cover with respect to the $\|\cdot\|_{P_n}$-norm. Since $f_{r_0(X)}$ is bounded and Lipschitz and μ_x is bounded and uniformly Lipschitz, the product $\mu_x \cdot f_{r_0(X)}$ is also uniformly Lipschitz. Therefore,

$$\begin{aligned}&\left|E_X\left[h_+^{-1} K\left(\frac{r_0(X) - r_0(X_i)}{h}\right)\delta_x(X) - f_0(X_i)\mu_x(r_0(X_i))\right]\right|\\ =&\left|\int K(t)\Big(\big(\mu_x \cdot f_{r_0(X)}\big)(r_0(X_i) + ht) - \big(\mu_x \cdot f_{r_0(X)}\big)(r_0(X_i))\Big)\,dt\right| \le C\,h\end{aligned}$$

for a constant C that is independent of x and X_i. This insight can be used to bound the empirical diameter

$$\widehat{\operatorname{diam}}_n = \sup_{x\in\mathbb{R}^{d_x}} \left\|G_{a,i}(x) - G_{b,i}(x)\right\|_{P_n} \le C\,h\left(\frac{1}{n}\sum_i |\epsilon_i|^2\right)^{\frac{1}{2}} \le C\,h.$$

Note that

$$\begin{aligned}\int_0^{\widehat{\operatorname{diam}}_n} \sqrt{\log(u^{-1} \wedge 1)}\,du \le& \int_0^{\widehat{\operatorname{diam}}_n \wedge 1} \sqrt{\log(u^{-1})}\,du\\ \le& \int_0^{\widehat{\operatorname{diam}}_n \wedge 1} 1 + \log(u^{-1})\,du\\ \le&(\widehat{\operatorname{diam}}_n \wedge 1) + (\widehat{\operatorname{diam}}_n \wedge 1)^2 \le 2(\widehat{\operatorname{diam}}_n \wedge 1)\log n,\end{aligned}$$

where the third inequality is due to $\int_0^x \log(u^{-1})\,du = x[x + \log(x^{-1})] \le x^2$ for $0 \le x \le 1$. Therefore, for large $\tilde{M}$, the entropy integral can be bounded by

$$\int_0^{\widehat{\operatorname{diam}}_n} \sqrt{\log \mathcal{N}\left(u, \mathcal{G}, \|\cdot\|_{P_n}\right)} \le \tilde{C}\int_0^{\widehat{\operatorname{diam}}_n} \sqrt{H_n(u)} \le 3\tilde{C}\,\widehat{\operatorname{diam}}_n\sqrt{\log n} \le \tilde{M} h \log n.$$

Lemma 5.1 in van de Geer 2000 in conjunction with Corollary 3.4 from the same book gives the exponential inequality

$$P\left(A_n \wedge \sup_{x\in\mathbb{R}^{d_x}} \frac{1}{n}\left|\sum_i \big(G_{a,i}(x) - G_{b,i}(x)\big)\right| \ge \tilde{M}\,n^{-\frac{1}{2}}h\log n\right) \le C\exp\left[-\frac{\tilde{M}^2(h\log n)^2}{64C^2\widehat{\operatorname{diam}}_n^2}\right].$$

The conclusion follows by noting that the right-hand side of the exponential inequality vanishes. □

Proof of Theorem 2.2 For K^* given in Lemma 2.1 let

$$G^{\alpha}_{0,ij}(r,x) = \delta_x(X_i)\,K_{h,ij}(r)\,\frac{h^{\alpha}}{\alpha!}\,\partial^{\alpha} m(r_0(X_j))\mathbf{1}_{\{K^*_{h,ij}>0\}}\left(\frac{r_0(X_i) - r_0(X_j)}{h}\right)^{\alpha}$$

and $P^\alpha_{0,n}(r,x) = \frac{1}{n(n-1)}\sum_{i\neq j} G^\alpha_{0,ij}(r,x)$. Note that due to $\left|K_{h,ij}(r)\right| \le K^*_{h,ij}(r_0)$ for $r \in \mathcal{R}_n$ the indicator function in this definition can be dropped without changing the expression. For a function $\tilde r$ that maps pairs (x_1, x_2) onto a vector of points on the line segment connecting $r_0(x_1)$ and $r_0(x_2)$ let

$$\tilde G^\alpha_{0,ij}(r,x) = \delta_x(X_i)K_{h,ij}(r)\frac{1}{(q+1)!}h^\alpha \partial^\alpha m(\tilde r(X_i,X_j))\mathbf{1}_{\{K^*_{h,ij}>0\}}\left(\frac{r_0(X_i)-r_0(X_j)}{h}\right)^\alpha.$$

The intermediate value function $\tilde r$ can be chosen such that

$$\begin{aligned}&\frac{1}{n(n-1)}\sum_{i\neq j}\delta_x(X_i)K_{h,ij}(\hat r)\big[m(r_0(X_i)) - m(r_0(X_j))\big]\\ =&\frac{1}{n(n-1)}\sum_{i\neq j}\left(\sum_{1\le|\alpha|\le q} G^\alpha_{0,ij}(r,x) + \sum_{|\alpha|=q+1}\tilde G^\alpha_{0,ij}(r,x)\right) = \sum_{1\le|\alpha|\le q} P^\alpha_{0,n}(r,x) + R_n(r,x).\end{aligned}$$

Below, it is shown that $\sup_{r\in\mathcal{R}_n, x\in\mathbb{R}^{d_x}}|R_n(r,x)| = o_p\big(n^{-\frac{1}{2}}\big)$. The remainder of the proof is based on the following stochastic decomposition

$$\begin{aligned}P^\alpha_{0,n}(\hat r,x) =& \big\{P^\alpha_{0,n}(\hat r,x) - P^\alpha_{0,n}(r_0,x) - E\big[P^\alpha_{0,n}(\hat r,x) - P^\alpha_{0,n}(r_0,x)\big]\big\}\\ &+\big\{P^\alpha_{0,n}(r_0,x) - E\,P^\alpha_{0,n}(r_0,x)\big\} + E\,P^\alpha_{0,n}(\hat r,x)\\ =&A_1 + A_2 + E\,P^\alpha_{0,n}(\hat r,x).\end{aligned}$$

In the sequel I show that the terms A_1 and A_2 vanish at the parametric rate. The expansion given in the theorem follows then from an expansion of $E\,P^\alpha_{0,n}(\hat r,x)$. To characterize the behavior of A_1 define

$$d^\alpha_{0,ij}(r,x) = n^{\eta_{\min}}\big(G^\alpha_{0,ij}(r,x) - G^\alpha_{0,ij}(r_0,x)\big).$$

Applying Lemma 2.5 with

$$f_n(W_i,W_j) = \frac{n^{-\eta_{\min}}h^\alpha}{\alpha!}\partial^\alpha m(r_0(X_i))\mathbf{1}_{\{K^*_{h,ij}>0\}}\left(\frac{r_0(X_j)-r_0(X_i)}{h}\right)^\alpha$$

yields

$$\begin{aligned}\sup_{r\in\mathcal{R}_n,x\in\mathbb{R}^{d_x}}\Bigg|\frac{1}{n(n-1)}\sum_{i\neq j}\Big(d^\alpha_{0,ij}(r,x)-&E_{W_i}d^\alpha_{0,ij}(r,x)\\ &-E_{W_j}d^\alpha_{0,ij}(r,x) + E\,d^\alpha_{0,ij}(r,x)\Big)\Bigg| &= o_p\big(n^{-\lambda_1}\big)\end{aligned}$$

for all

$$\lambda_1 < \frac{1}{2} + \frac{1}{2}\big(1-\eta_+\big) + (\delta-\eta)_{\min} - \max_{1\le k\le d}\left[\delta_k\gamma_k + \xi_k\right].$$

By Lemma 2.2

$$\begin{aligned}\sup_{r\in\mathcal{R}_n,x\in\mathbb{R}^{d_x}}\left|\frac{1}{n(n-1)}\sum_{i\neq j}\Big(E_{W_i}d^\alpha_{0,ij}(r,x) - E\,d^\alpha_{0,ij}(r,x)\Big)\right| = o_p\big(n^{-\lambda_2}\big)\\ \sup_{r\in\mathcal{R}_n,x\in\mathbb{R}^{d_x}}\left|\frac{1}{n(n-1)}\sum_{i\neq j}\Big(E_{W_j}d^\alpha_{0,ij}(r,x) - E\,d^\alpha_{0,ij}(r,x)\Big)\right| = o_p\big(n^{-\lambda_2}\big)\end{aligned}$$

for all

$$\lambda_2 < \frac{1}{2} + (\delta - \eta)_{\min} - \frac{1}{2} \max_{1 \le k \le d} [\delta_k \gamma_k + \xi_k] .$$

Collecting these results and noting that

$$P^{\alpha}_{0,n}(\hat{r}, x) - P^{\alpha}_{0,n}(r_0, x) = n^{-\eta_{\min}} \frac{1}{n(n-1)} \sum_{i \ne j} d^{\alpha}_{0,ij}(\hat{r}, x)$$

gives

$$\sup_{x \in \mathbb{R}^{d_x}} |A_1| = o_p\left(n^{-(\lambda_1 + \eta_{\min})} + n^{-(\lambda_1 + \eta_{\min})}\right) = o_p\big(n^{-\frac{1}{2}}\big).$$

Repeating similar arguments for $|\alpha| = q + 1$ and

$$\tilde{d}^{\alpha}_{0,ij}(r, x) = n^{(q+1)\eta_{\min}} \big(\tilde{G}^{\alpha}_{0,ij}(r, x) - \tilde{G}^{\alpha}_{0,ij}(r_0, x)\big)$$

gives

$$\sup_{r \in \mathcal{R}_n, x \in \mathbb{R}^{d_x}} \left| \frac{1}{n(n-1)} \sum_{i \ne j} \tilde{G}^{\alpha}_{0,ij}(r, x) - E\, \tilde{G}^{\alpha}_{0,ij}(r, x) \right| = o_p\big(n^{-\frac{1}{2}}\big).$$

There is a universal constant $C > 0$ such that $\left|E\, \tilde{G}^{\alpha}_{0,ij}(r, x)\right| \le C\, n^{-(q+1)\eta_{\min}}$ and hence

$$\sup_{r \in \mathcal{R}_n, x \in \mathbb{R}^{d_x}} |R_n(r, x)| = \sup_{r \in \mathcal{R}_n, x \in \mathbb{R}^{d_x}} \left| \frac{1}{n(n-1)} \sum_{i \ne j} \sum_{|\alpha| = q+1} \tilde{G}^{\alpha}_{0,ij}(r, x) \right| = o_p\big(n^{-\frac{1}{2}}\big).$$

By Lemma 2.7

$$\sup_{x \in \mathbb{R}^{d_x}} |A_2| = o_p\big(n^{-\frac{1}{2} - \eta_{\min}}\big) = o_p\big(n^{-\frac{1}{2}}\big).$$

Expanding $E\, P^{\alpha}_{0,n}(x, \hat{r})$ gives

$$\begin{aligned}
&E\, P^{\alpha}_{0,n}(\hat{r}, x) \\
=&E\, P^{\alpha}_{n}(\hat{r}, x) - \Delta(\hat{r}, x) \\
&- E\Bigg[\delta_x(X_i) K_{h,ij}(r_0) \sum_{q^*+1 \le |\alpha| \le q} \frac{h^{\alpha}}{\alpha!} \partial^{\alpha} m(r_0(X_i)) \Bigg\{ \left(\frac{\hat{r}(X_j) - \hat{r}(X_i)}{h}\right)^{\alpha} \\
&\qquad\qquad - \left(\frac{r_0(X_j) - r_0(X_i)}{h}\right)^{\alpha} \Bigg\}\Bigg] \\
&- E\Bigg[\delta_x(X_i)\big(K_{h,ij}(\hat{r}) - K_{h,ij}(r_0)\big) \\
&\qquad \sum_{1 \le |\alpha| \le q} \frac{h^{\alpha}}{\alpha!} \partial^{\alpha} m(r_0(X_j)) \Bigg\{ \left(\frac{\hat{r}(X_j) - \hat{r}(X_i)}{h}\right)^{\alpha} - \left(\frac{r_0(X_j) - r_0(X_i)}{h}\right)^{\alpha} \Bigg\}\Bigg] \\
=&E\, P^{\alpha}_{n}(\hat{r}, x) - \Delta(\hat{r}, x) + A_{3a} + A_{3b}.
\end{aligned}$$

Since $\hat{r} \in \mathcal{R}_n$ we have

$$\sup_{x \in \mathbb{R}^{d_x}} |A_{3a}| = o_p\Big(n^{-((\delta-\eta)_{\min}+(q^*+1)\eta_{\min})}\Big) = o_p\big(n^{-\frac{1}{2}}\big).$$

By Lemma 2.1 there is an integrable K^* such that

$$\big|K_{h,ij}(\hat{r}) - K_{h,ij}(r_0)\big| \le K^*_{h,ij}(r_0) n^{-(\delta-\eta)_{\min}}.$$

Hence

$$\sup_{x \in \mathbb{R}^{d_x}} |A_{3b}| = o_p\Big(n^{-(\delta_{\min}+(\delta-\eta)_{\min})}\Big) = o_p\big(n^{-\frac{1}{2}}\big). \qquad \square$$

Proof of Theorem 2.3 Let $\mathcal{D} = \{x \in \mathbb{R}^{d_x} : \|r_0(x) - \mathrm{supp}(r_0(X))\| \le 1\}$ and let

$$\tilde{K}^{\alpha}_{h,ij}(r) = K_{h,ij}(r)\left(\frac{r(X_j) - r(X_i)}{h}\right)^{\alpha}.$$

Let $D^{\alpha}_n(x) = \widehat{\partial^{\alpha} m}(\hat{r}(x)) - \partial^{\alpha} m(r_0(x))$. Under the assumptions of the theorem, this term can be bounded uniformly. By Lipschitz continuity of all derivatives of m up to qth order

$$\begin{aligned}\sup_{x\in\mathcal{D}} |D^{\alpha}_n(x)| \le& \big|\widehat{\partial^{\alpha} m}(\hat{r}(x)) - \partial^{\alpha} m(\hat{r}(x))\big| + \big|\partial^{\alpha} m(\hat{r}(x)) - \partial^{\alpha} m(r_0(x))\big| \\ \le& \sup_t \big|\widehat{\partial^{\alpha} m}(t) - \partial^{\alpha} m(t)\big| + C \max_{j=1,\dots,d} \sup_{x\in\mathcal{D}} \big|\hat{r}_j(x) - r_{0,j}(x)\big| \\ \le& O_p\Big(n^{-\theta_{|\alpha|}} + n^{-\delta_{\min}}\Big).\end{aligned}$$

The proof is based on the following stochastic decomposition of the estimator

$$\begin{aligned}\hat{P}^{\alpha}_n(x) =& \frac{h^{\alpha}}{\alpha!}\frac{1}{n(n-1)}\sum_{i\neq j}\Big\{\delta_x(X_i) D^{\alpha}_n(x) \tilde{K}^{\alpha}_{h,ij}(r_0)\Big\} \\ &+ \frac{h^{\alpha}}{\alpha!}\frac{1}{n(n-1)}\sum_{i\neq j}\Big\{\delta_x(X_i) D^{\alpha}_n(x)\big(\tilde{K}^{\alpha}_{h,ij}(\hat{r}) - \tilde{K}^{\alpha}_{h,ij}(r_0)\big)\Big\} \\ &+ \big(P^{\alpha}_n(x, r_0) - E\, P^{\alpha}_n(x, r_0)\big) \\ &+ \big(P^{\alpha}_n(x, \hat{r}) - P^{\alpha}_n(x, r_0) + E\, P^{\alpha}_n(x, r_0)\big) = I_1 + I_2 + I_3 + I_4.\end{aligned}$$

The strategy of the proof is to show that the terms I_1 through I_3 vanish at the desired rate. Finally, it is shown that I_4 yields the desired limit. By standard arguments

$$\begin{aligned}\sup_{x\in\mathcal{D}} h_+^{-1}\Bigg|&\frac{1}{n}\sum_j K\left(\frac{r_0(x) - r_0(X_j)}{h}\right)\left(\frac{r_0(X_j) - r_0(x)}{h}\right)^{\alpha} \\ &- E\, K\left(\frac{r_0(x) - r_0(X_j)}{h}\right)\left(\frac{r_0(X_j) - r_0(x)}{h}\right)^{\alpha}\Bigg| = O_p\left(\sqrt{\frac{\log n}{n h_+}}\right).\end{aligned}$$

Moreover, for an appropriate intermediate value function $\tilde{y}$

$$\begin{aligned}
&E\,K\left(\frac{r_0(x)-r_0(X_j)}{h}\right)\left(\frac{r_0(X_j)-r_0(x)}{h}\right)^{\alpha}\\
=&\int h_+^{-1}K\left(\frac{r_0(x)-y}{h}\right)\left(\frac{y-r_0(x)}{h}\right)^{\alpha} f_{r_0(X)}(y)\,dy\\
=&(-1)^{\alpha}\int\left\{K(y)y^{\alpha}\left(f_{r_0(X)}(r_0(x))+\sum_{1\le|\beta|\le q_1\wedge q_2}\partial^{\beta}f_{r_0(X)}(r_0(x))(-hy)^{\beta}\right)\right\}dy\\
&+(-1)^{\alpha}\int K(y)y^{\alpha}\sum_{|\beta|=q_1\wedge q_2+1}\partial^{\beta}f_{r_0(X)}(\tilde{y}(x,y))(-hy)^{\beta}\,dy.
\end{aligned}$$

The first term on the right-hand side is zero and the second term on the right-hand side can be bounded by $C\,n^{-(q_1\wedge q_2+1)\eta_{\min}}$, where the choice of the constant C does not depend on x. Therefore,

$$\begin{aligned}
&\sup_{x\in\mathcal{D}}h_+^{-1}\left|\frac{1}{n}\sum_j K\left(\frac{r_0(x)-r_0(X_j)}{h}\right)\left(\frac{r_0(X_j)-r_0(x)}{h}\right)^{\alpha}\right|\\
&\qquad=O_p\left(\sqrt{\frac{\log n}{nh_+}}+n^{-(q_1\wedge q_2+1)\eta_{\min}}\right).
\end{aligned}$$

Collecting the results so far gives

$$\begin{aligned}
I_1=&h^{\alpha}\frac{1}{n}\sum_i\left(\delta_x(X_i)D_n^{\alpha}(x)\frac{1}{n-1}\sum_{j:j\neq i}\tilde{K}_{h,ij}^{\alpha}(r_0)\right)\\
=&n^{-|\alpha|\eta_{\min}}O_p\left(\left(n^{-\theta_{|\alpha|}}+n^{-\delta_{\min}}\right)\left(\sqrt{\frac{\log n}{nh_+}}+n^{-(q_1\wedge q_2+1)\eta_{\min}}\right)\right)=o_p\left(n^{-\kappa}\right).
\end{aligned}$$

Using Lemma 2.1 to bound the summands in I_2 gives

$$I_2\le C\,n^{-|\alpha|\eta_{\min}-(\delta-\eta)_{\min}}O_p\left(n^{-\theta_{|\alpha|}}+n^{-\delta_{\min}}\right)=o_p\left(n^{-\kappa}\right).$$

By Lemma 2.7

$$I_3=o_p\left(n^{-\kappa_7}\right).$$

To tackle the I_4 term let

$$\Delta_{ij}^{\alpha}(r,x)=n^{\eta_{\min}}\frac{h^{\alpha}}{\alpha!}\partial^{\alpha}m(r_0(X_i))\left(\tilde{K}_{h,ij}^{\alpha}(r)-\tilde{K}_{h,ij}^{\alpha}(r_0)\right).$$

Lemma 2.5 implies

$$\begin{aligned}
\sup_{r\in\mathcal{R}_n,x\in\mathbb{R}^{d_x}}&\left|\sum_{i\neq j}\frac{1}{n(n-1)}\left(\Delta_{ij}^{\alpha}(r,x)-E_{W_i}\Delta_{ij}^{\alpha}(r,x)\right.\right.\\
&\qquad\left.\left.-E_{W_j}\Delta_{ij}^{\alpha}(r,x)+E\,\Delta_{ij}^{\alpha}(r,x)\right)\right|=o_p\left(n^{-(\kappa_5-\eta_{\min})}\right).
\end{aligned}$$

Now, Lemma 2.2 gives

$$\begin{aligned}\sup_{r\in\mathcal{R}_n, x\in\mathbb{R}^{d_x}} \left|\frac{1}{n(n-1)}\sum_{i\neq j}\left(E_{W_i}\Delta_{ij}^{\alpha}(r,x) - E\,\Delta_{ij}^{\alpha}(r,x)\right)\right| &= o_p\left(n^{-(\kappa_6-\eta_{\min})}\right)\\ \sup_{r\in\mathcal{R}_n, x\in\mathbb{R}^{d_x}} \left|\frac{1}{n(n-1)}\sum_{i\neq j}\left(E_{W_j}\Delta_{ij}^{\alpha}(r,x) - E\,\Delta_{ij}^{\alpha}(r,x)\right)\right| &= o_p\left(n^{-(\kappa_6-\eta_{\min})}\right).\end{aligned}$$

Since we can write $P_n^{\alpha}(x,\hat{r}) - P_n^{\alpha}(x,r_0) = n^{-\eta_{\min}}\frac{1}{n(n-1)}\sum_{i\neq j}\Delta_{ij}^{\alpha}(x,\hat{r})$ this implies

$$\begin{aligned}&\sup_{x\in\mathbb{R}^{d_x}}\left|P_n^{\alpha}(x,\hat{r}) - P_n^{\alpha}(x,r_0) - \left(E\,P_n^{\alpha}(x,\hat{r}) - E\,P_n^{\alpha}(x,r_0)\right)\right|\\ &\qquad = o_p\left(n^{-\kappa_5} + n^{-\kappa_6}\right) = o_p\left(n^{-\kappa}\right).\end{aligned}$$

Plugging I_4 into this equation yields

$$\sup_{x\in\mathbb{R}^{d_x}}\left|I_4 - E\,P_n^{\alpha}(x,\hat{r})\right| = o_p\left(n^{-\kappa}\right)$$

and hence the conclusion. □

Proof of Theorem 2.4 Appendix 2.D gives a uniform expansion of the local polynomial estimator.

$$\begin{aligned}\hat{r}(x^{(1)}) =& r_0(x^{(1)}) + e_1'[E\mathbf{S}_n(x^{(1)})]^{-1}\frac{1}{n}\sum_i \epsilon_i\,\Gamma\left(\frac{X_i^{(1)} - x^{(1)}}{g}\right)L_g(X_i^{(1)} - x^{(1)})\\ &+ e_1'B_n^*(x^{(1)}) + \tilde{R}_n(x^{(1)})\\ =& \tilde{r}_n(x^{(1)}) + \tilde{R}_n(x^{(1)})\end{aligned}$$

with

$$\sup_{x^{(1)}\in\mathcal{D}}\left|\tilde{R}_n(x^{(1)})\right| = O_p\left(\frac{\log n}{ng_+} + n^{-(p+1)\eta_{\min}^*}\sqrt{\frac{\log n}{ng_+}}\right)$$

First, I show that we can take $\hat{r} = \tilde{r}$ at the expense of a $o_p\left(n^{-\frac{1}{2}}\right)$-term.

$$\begin{aligned}&\left|\frac{1}{n(n-1)}\sum_{i\neq j}\delta_x(X_i)h_+^{-1}\left[K\left(\frac{\hat{r}(X_i) - \hat{r}(X_j)}{h}\right) - K\left(\frac{\tilde{r}(X_i) - \tilde{r}(X_j)}{h}\right)\right]\epsilon_i\right|\\ \leq& C\sum_{k=1}^{d}\frac{\left\|\tilde{R}_{k,n}\right\|_{\infty}}{h_k}\left\{\frac{1}{n(n-1)}\sum_{i\neq j}h_+^{-1}K^*\left(\frac{r_0(X_i) - r_0(X_j)}{h}\right)|\epsilon_i|\right\}\\ \leq& C\,(\log n)n^{-\frac{1}{2}}\min_{k=1,\ldots,d}\left(n^{-\left(\frac{1}{2}-d_{x,k}^{(1)}\eta_k^*-\eta_k\right)} + n^{-\left((p+1)\eta_k^*-\frac{1}{2}d_{x,k}^{(1)}\eta_k^*-\eta_k\right)}\right)O_p\left(n^{-\frac{1}{2}(1-\eta_+)} + 1\right)\\ =& o_p\left(n^{-\frac{1}{2}}\right).\end{aligned}$$

For the other terms in $\hat{T}_n$ argue similarly. The approach for the rest of the proof is to apply Theorem 2.1 and Theorem 2.2 with $\tilde{r}$ replacing $\hat{r}$ and $q^* = 1$. According to Appendix 2.D we can take

$$\eta = \eta^* = \frac{1}{5}, \qquad \delta = \frac{2}{5}, \qquad \gamma = \frac{1}{2} \quad \text{and} \qquad \xi > 0$$

when checking the assumptions of Theorem 2.1 and Theorem 2.2. It is straightforward to verify that all restrictions are met. Therefore, uniformly in x

$$\begin{aligned} \hat{T}_{\text{error},n}(x) =& \frac{1}{n}\sum_i f_{r_0(X)}(r_0(X_i))\big(\delta_x(X_i) - \mu_x(r_0(X_i))\big)\epsilon_i + o_p\big(n^{-\frac{1}{2}}\big). \\ T_{\text{bias, n}}(x) =& -\Delta(\tilde{r}, x) + \Xi(\tilde{r}, x) + o_p\big(n^{-\frac{1}{2}}\big). \end{aligned}$$

Next, I characterize $\Delta(\tilde{r}, x)$. The goal is to show that

$$\sup_{x \in \mathbb{R}^{d_x}} \big|\Delta(\tilde{r}, x) - \tilde{\Delta}(x)\big| = o_p\big(n^{-\frac{1}{2}}\big).$$

where

$$\tilde{\Delta}(x) = -\frac{1}{n}\sum_i \zeta_i\big[\delta_x(X_i) - \mu_x(r_0(X_i))\big]\partial m(r_0(X_i)) f_{r_0(X)}(r_0(X_i)).$$

Split $\tilde{r}$ into a bias and an error part. Start with the bias part

$$E\Big[\delta_x(X_i) K_{h,ij}(r_0)\partial m(r_0(X_i))\big\{B_n^*(X_j^{(1)}) - B_n^*(X_i^{(1)})\big\}\Big].$$

Note that

$$E_X h_+^{-1} K\left(\frac{r_0(X_i) - r_0(X)}{h}\right) = f_{r_0(X)}(r_0(X_i)) + O(\|h\|)$$

and uniformly in x

$$\begin{aligned} E_X \mu_x(r_0(X)) h_+^{-1} K\left(\frac{r_0(X) - r_0(X_j)}{h}\right) \partial m(r_0(X)) \\ = \mu_x(r_0(X_j))\partial m(r_0(X_j)) f_{r_0(X)}(r_0(X_j)) + O(\|h\|) \end{aligned}$$

Therefore, uniformly in x

$$\begin{aligned} & E\Big[\delta_x(X_i) K_{h,ij}(r_0)\partial m(r_0(X_i)) B_n^*(X_i^{(1)})\Big] \\ =& E\Big[\delta_x(X_i)\partial m(r_0(X_i)) f_{r_0(X)}(r_0(X_i)) B_n^*(X_i^{(1)})\Big] + O\Big(n^{-(p+1)\eta^* - \eta}\Big) \qquad \text{and} \\ & E\Big[\delta_x(X_i) K_{h,ij}(r_0)\partial m(r_0(X_i)) B_n^*(X_j^{(1)})\Big] \\ =& E\Big[\mu_x(X_j)\partial m(r_0(X_j)) f_{r_0(X)}(r_0(X_j)) B_n^*(X_j^{(1)})\Big] + O\Big(n^{-(p+1)\eta^* - \eta}\Big). \end{aligned}$$

Hence, uniformly in x

$$\begin{aligned}&E\Big[\delta_x(X_i)K_{h,ij}(r_0)\partial m(r_0(X_i))\big\{B_n^*(X_i^{(1)})-B_n^*(X_j^{(1)})\big\}\Big]\\ =&E\Big[\big\{\delta_x(X_i)-\mu_x(X_i)\big\}K_{h,ij}(r_0)\partial m(r_0(X_i))B_n^*(X_i^{(1)})\Big]+o\Big(n^{-\frac{1}{2}}\Big).\end{aligned}$$

Next, let

$$\varphi_n(x^{(1)})=\frac{1}{n}\sum_i\zeta_i\,\Gamma\left(\frac{X_i^{(1)}-x^{(1)}}{g}\right)L_g(X_i^{(1)}-x^{(1)}).$$

and turn to the error part

$$E\Big[\delta_x(X_i)K_{h,ij}(r_0)\partial m(r_0(X_i))e_1'\big\{[E\mathbf{S}_n(X_j^{(1)})]^{-1}\varphi_n(X_j^{(1)})-[E\mathbf{S}_n(X_i^{(1)})]^{-1}\varphi_n(X_i^{(1)})\big\}\Big].$$

Let $\nu_1(u)=E[K_{h,ij}(r_0)\mid r_0(X_i)=u]$. Uniformly in u

$$\nu_1(u)=f_{r_0(X)}(u)+O\big(\,\|h\|\,\big).$$

For vectors u and g of the same dimension write $u_g=(\operatorname{Diag}g)u$ and let

$$\mathbf{M}=\int\Gamma(u)\Gamma'(u)L(u)\,du.$$

Note that

$$\begin{aligned}&\sup_{v\in\mathbb{R}^{d_x^{(1)}}}\left|E\mathbf{S}_n(v-u_g)f_{X^{(1)}}^{-1}(v-u_g)-\mathbf{M}\right|\\ =&\sup_{v\in\mathbb{R}^{d_x^{(1)}}}\left|\int\Gamma(t)\Gamma'(t)L(t)\left\{\frac{f_{X^{(1)}}(v-u_g+t_g)}{f_{X^{(1)}}(v-u_g)}-1\right\}dt\right|=O(\|g\|)\end{aligned}$$

Therefore, uniformly in $v\in\mathcal{D}$

$$\begin{aligned}&\int\delta_x(u)\nu_1(r_0(u))\partial m(r_0(u))e_1'\big[E\mathbf{S}_n(u)f_{X^{(1)}}^{-1}(u)\big]^{-1}\Gamma\left(\frac{v-u}{g}\right)g_+^{-1}L\left(\frac{v-u}{g}\right)du\\ =&\int\delta_x(v-u_g)\nu_1(r_0(v-u_g))\partial m(r_0(v-u_g))\\ &\qquad e_1'\big[E\mathbf{S}_n(v-u_g)f_{X^{(1)}}^{-1}(v-u_g)\big]^{-1}\Gamma\,(u)\,L\,(u)\;du\\ =&\int\delta_x(v-u_g)\big\{f_{r_0(X)}(r_0(v))+O(\|g\|)+O(\|h\|)\big\}\big\{\partial m(r_0(v))+O(\|g\|)\big\}\\ &\qquad\big\{e_1'\mathbf{M}^{-1}+O(\|g\|)\big\}\Gamma(u)L(u)\,du\\ =&\,\delta_x(v)f_{r_0(X)}(r_0(v))\partial m(r_0(v))+\\ &\qquad f_{r_0(X)}(r_0(v))\partial m(r_0(v))\int(\delta_x(v-u_g)-\delta_x(v))e_1'\mathbf{M}^{-1}\Gamma(u)L(u)\,du\\ &+O(\|g\|+\|h\|)\\ =&\,\delta_x(v)f_{r_0(X)}(r_0(v))\partial m(r_0(v))+O(\|g\|+\|h\|).\end{aligned}$$

Hence, arguing similarly to the proof of Theorem 2.1 gives

$$\sup_{x\in\mathbb{R}^{d_x}} \Big| E\big\{\delta_x(X_i)K_{h,ij}(r_0)\partial m(r_0(X_i))e_1'\big[E\mathbf{S}_n(X_i)\big]^{-1}\varphi_n(X_i)\big\} \\ - \frac{1}{n}\sum_i \zeta_i\big[\delta_x(X_i)\partial m(r_0(X_i))f_{r_0(X)}(r_0(X_i))\big]\Big| = o_p\big(n^{-\frac{1}{2}}\big).$$

Let $\nu_2(X_j) = E_{X_i}[K_{h,ij}(r_0)\delta_x(X_i)\partial m(r_0(X_i))]$. Uniformly in u

$$\nu_2(u) = \mu_x(r_0(u))\partial m(r_0(u))f_{r_0(X)}(r_0(u)) + O(\|h\|).$$

Uniformly in v

$$\begin{aligned}
&\int \nu_2(u)e_1'\big[E\mathbf{S}_n(u)f^{-1}_{X^{(1)}}(u)\big]^{-1}\Gamma\left(\frac{v-u}{g}\right)g_+^{-1}L\left(\frac{v-u}{g}\right)du\\
&=\int \nu_2(v-u_g)e_1'\big[E\mathbf{S}_n(v-u_g)f^{-1}_{X^{(1)}}(v-u_g)\big]^{-1}\Gamma(u)L(u)\,du + O(\|h\|)\\
&=\nu_2(v)e_1'\mathbf{M}^{-1}\int \Gamma(u)L(u)\,du + O(\|g\|+\|h\|)\\
&=\mu_x(r_0(v))\partial m(r_0(v))f_{r_0(X)}(r_0(v)) + O(\|g\|+\|h\|).
\end{aligned}$$

Therefore,

$$\sup_{x\in\mathbb{R}^{d_x}} \Big| E\big\{\delta_x(X_i)K_{h,ij}(r_0)\partial m(r_0(X_i))e_1'\big[E\mathbf{S}_n(X_j^{(1)})\big]^{-1}\varphi_n(X_j^{(1)})\big\} \\ - \frac{1}{n}\sum_i \zeta_i\big[\mu_x(r_0(v))\partial m(r_0(v))f_{r_0(X)}(r_0(v))\big]\Big| = o_p\big(n^{-\frac{1}{2}}\big). \qquad \square$$

Appendix 2.C Lemmas

The following lemma is a variation of an argument in B. Hansen 2008.

Lemma 2.1 *Let $u_1 \in \mathbb{R}^d$ and $u_2 \in \mathbb{R}^d$ lie in an ϵ-ball around $u_0 \in \mathbb{R}^d$. Under Assumption 2.1 (ii)*

$$|K(u_1) - K(u_2)| \le K^*(u_0)\,\|u_1 - u_2\|$$

for a function K^* depending only only on K and ϵ. K^* can be chosen such that it is positive, bounded, has bounded support, is Lipschitz continuous and satisfies $|K| \le K^*$, $|K(u_1)| \le K^*(u_0)$, $|K(u_2)| \le K^*(u_0)$. Moreover, for $\alpha \in \mathbb{N}_0^d$ there is a constant C_α depending only on K, α and ϵ such that

$$|K(u_1)u_1^\alpha - K(u_2)u_2^\alpha| \le C_\alpha K^*(u_0)\,\|u_1 - u_2\|.$$

Proof There is a constant $\tilde{L}$ such that

$$|K(u_1) - K(u_2)| \le \tilde{L} \, \|u_1 - u_2\| \, .$$

$$\begin{aligned} |K(u_1) - K(u_2)| &\le |K(u_1)| + |K(u_2)| \\ &\le L^d \Big(1_{\{\|u_1\| \le L\}} + 1_{\{\|u_2\| \le L\}} \Big) \\ &\le \big(2L^d + \tilde{L} \big) 1_{\{\|u_0\| \le L + \epsilon\}} \overset{\text{def}}{=} K^{**}(u_0). \end{aligned}$$

K^{**} satisfies all the properties required of K^* with the exception of Lipschitz continuity. Choose K^* to be an appropriate Lipschitz continuous majorant of K^{**}. For the second claim write

$$|K(u_1)u_1^\alpha - K(u_2)u_2^\alpha| \le |(K(u_1) - K(u_2))u_1^\alpha| + |K(u_2)(u_1^\alpha - u_2^\alpha)| = I_1 + I_2.$$

Since K^* has bounded support and since u_1 and u_2 are close to u_0, there is a constant C such that

$$I_1 \le C K^*(u_0) \, \|u_1 - u_2\| \, .$$

For I_2, note that u_1 and u_2 are bounded if u_0 is in the support of K^* so that

$$|u_1^\alpha - u_2^\alpha| \le C \, \mathbf{1}_{\{K^*(u_0) > 0\}} \max_{j=1,\ldots,d} \left| u_{1,j}^{\alpha_j} - u_{2,j}^{\alpha_j} \right|$$

for a constant C. For $j = 1, \ldots, d$ with $\alpha_j \ge 1$

$$\left| u_{1,j}^{\alpha_j} - u_{2,j}^{\alpha_j} \right| \le \alpha_j \, |\hat{u}|^{\alpha_j - 1} \, |u_{1,j} - u_{2,j}|$$

for $\hat{u}$ between $u_{1,j}$ and $u_{2,j}$. If u_0 is in the support of K^*, then $\hat{u}$ is bounded. Using $K(u_1) \le K^*(u_0)$ we have

$$I_2 \le C K^*(u_0) \max_{j=1,\ldots,d} |u_{1,j} - u_{2,j}| \le C K^*(u_0) \, \|u_1 - u_2\| \, .$$

□

Lemma 2.2 *Let Φ and Ψ denote bounded VC-classes. Let $\Theta = \Phi \times \Psi \times \mathcal{R}_n$ and let $\theta = (\phi, \tau, r)$ denote a generic element from Θ. For any $\theta = (\phi, \psi, r) \in \Theta$ let $\theta_0 = (\phi, \psi, r_0)$. Let $f_n : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote a function satisfying $\limsup_{n\to\infty} \sup_{w_1 \in \mathcal{W}, w_2 \in \mathcal{W}} |f_n(w_1, w_2)| < \infty$ and let $a_n : \mathcal{X} \to \mathbb{R}$ and $b_n : \mathcal{W} \to \mathbb{R}$ denote functions satisfying*

$$\limsup_{n\to\infty} E \, |a_n(X)|^s < \infty \quad \textit{and} \quad \limsup_{n\to\infty} E \, |b_n(X, \epsilon)|^t < \infty$$

for $s > 2$ and $t > 2$. For $\alpha \in \mathbb{N}_0^d$ define

$$G(\theta, x, u) = h_+^{-1} E_X \Bigg[\left(\frac{r(X) - r(x)}{h} \right)^\alpha K \left(\frac{r(X) - r(x)}{h} \right) \\ f_n(x, X) a_n(X) \phi(X) \Bigg] b_n(x, u) \psi(x)$$

and write $G_i(\theta) = G(\theta, X_i, \epsilon_i)$*. Fix a* $\kappa^* > 0$ *such that*

$$\kappa^* < \frac{1}{2} - \frac{\eta_+}{s} + (\delta - \eta)_{\min} - \frac{1}{2} \max_{1 \le j \le d} \big[\delta_j \gamma_j + \xi_j\big].$$

Then

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \Big(G_i(\theta) - G_i(\theta_0) - E\Big[G_i(\theta) - G_i(\theta_0)\Big]\Big) \right| = O_p\left(n^{-\kappa^*}\right).$$

Proof Let $\mathcal{G} = \{G(\theta) - G(\theta_0) : \theta \in \Theta\}$. First show that there are positive constants $\tilde{C}$ and $\tilde{u}$ such that on a set wpa1 for all $0 < u < \tilde{u}$

$$\log \mathcal{N}(u, \mathcal{G}, \|\cdot\|_{P_n}) \le \tilde{C}\, H_n(u),$$

with

$$H_n(u) = \log n + \sum_{j=1}^{d} n^{\xi_j + \gamma_j(\eta_j + \eta_+/s)} u^{-\gamma_j}.$$

Let $A_n = \{\|b_n\|_{P_n,s} \le C_A\}$ and choose C_A large enough that $PA_n \to 1$. Now work conditional on A_n. Construct a u-cover of $\mathcal{G}$ in the following way. Take $u_{\phi,\psi} = (C^*)^{-1} h_+(u/6)$ and for $j = 1, \dots, d$ take $u_{r,j} = (C^*)^{-1} n^{-\eta_+/s} h_j(u/6)$. Now, let $\mathcal{C}_n^\phi$ denote a $u_{\phi,\psi}$-cover of Φ with respect to the $\|\cdot\|_{P,\frac{s}{s-1}}$-norm and let $\mathcal{C}_n^\psi$ denote a $u_{\phi,\psi}$-cover of Ψ with respect to the $\|\cdot\|_{P_n,Q}$-norm, $Q = \frac{2s}{s-2}$. Since Φ and Ψ are VC there exist constants A and V such that $\mathcal{C}_n^\phi$ and $\mathcal{C}_n^\psi$ can be chosen to contain less than $Au_{\phi,\psi}^{-V}$ functions each (Theorem 2.6.4 in van der Vaart and Wellner 1996). For each $j = 1, \dots, d$ let $\mathcal{C}_{n,j}^r$ denote a $u_{r,j}$-cover of $\mathcal{R}_{n,j}$ with respect to the $\|\cdot\|_\infty$-norm. By construction

$$\mathcal{C}_n^G = \left\{ G(\theta) - G(\theta_0) : \theta \in \mathcal{C}_n^\phi \times \mathcal{C}_n^\psi \times \prod_{j=1}^{d} \mathcal{C}_{n,j}^r \right\}$$

is a cover of $\mathcal{G}$ and can be chosen such that

$$\log \#\mathcal{C}_n^G \le M_1 + M_2 \log n + M_3 \log u^{-1} + M_4 \sum_{j=1}^{d} \left(n^{\xi_j + \gamma_j(\eta_j + \eta_+/s)} u^{-\gamma_j} \right)$$

for constants $M_1, \dots, M_4$. It is wlog to assume $\xi_j > 0$ and $\gamma_j > 0$ for all $j = 1, \dots, d$. By l'Hôpital's rule

$$\lim_{u \to 0} \frac{\log u^{-1}}{u^{-\gamma_j}} = \infty$$

for all j. Thus, there is a $\tilde{u} > 0$ and a constant $\tilde{C}$ such that eventually

$$\log \#\mathcal{C}_n^G \le \tilde{C}\, H_n(u)$$

for all $u < \tilde{u}$. Next, show that if C^* is chosen large enough then $\mathcal{C}_n^G$ will be a u-cover of $\mathcal{G}$ with respect to the $\|\cdot\|_{P_n}$-norm. Let $\theta_1 = (\phi_1, \psi_1, r_1)$ denote a point in Θ and let $\theta_2 =$

(ϕ_2, ψ_2, r_2) denote a nearest grid-point, i.e., $\|\phi_1 - \phi_2\|_{P,\frac{s}{s-1}} \leq u_{\phi,\psi}$, $\|\psi_1 - \psi_2\|_{P_n,Q} \leq u_{\phi,\psi}$ and $\|r_{1,j} - r_{2,j}\|_\infty \leq u_{r,j}$ for $j = 1, \ldots, d$. To prove

$$\|G(\theta_1) - G(\theta_{1,0}) - [G(\theta_2) - G(\theta_{2,0})]\|_{P_n} \leq u$$

it suffices to show

$$\|G(\theta_1) - G(\theta_2)\|_{P_n} \leq \frac{u}{2}.$$

Decompose

$$\begin{aligned}\|G(\theta_1) - G(\theta_2)\|_{P_n} \leq & \|G(\phi_1, \psi_1, r_1) - G(\phi_2, \psi_1, r_1)\|_{P_n} \\ & + \|G(\phi_2, \psi_1, r_1) - G(\phi_2, \psi_2, r_1)\|_{P_n} \\ & + \|G(\phi_2, \psi_2, r_1) - G(\phi_2, \psi_2, r_2)\|_{P_n} \leq I_1 + I_2 + I_3.\end{aligned}$$

To bound the right-hand side note that on A_n

$$\left(\frac{1}{n}\sum_{i=1}^n |b_n(X_i, \epsilon_i)|^2\right)^{\frac{1}{2}} \leq \left(\frac{1}{n}\sum_{i=1}^n |b_n(X_i, \epsilon_i)|^t\right)^{\frac{1}{t}} \leq C_A.$$

To bound I_1 note that

$$\begin{aligned}& E_X \left|f_n(x, X) a_n(X) \big(\phi_1(X) - \phi_1(X)\big)\right| \\ \leq & C \left(E_X |a_n(X)|^s\right)^{\frac{1}{s}} \left(E_X |\phi_1(X) - \phi_2(X)|^{\frac{s}{s-1}}\right)^{\frac{s-1}{s}}.\end{aligned}$$

Therefore, for C^* chosen large enough

$$\begin{aligned} I_1 \leq & C\, h_+^{-1} \Big(E_X |\phi_1(X) - \phi_2(X)|^{\frac{s}{s-1}}\Big)^{\frac{s-1}{s}} \left(\frac{1}{n}\sum_{i=1}^n |b_n(X_i, \epsilon_i)|^2\right)^{\frac{1}{2}} \\ \leq & C^*\, h_+^{-1} \|\phi_1 - \phi_2\|_{P,\frac{s}{s-1}}.\end{aligned}$$

Since $E_X |f_n(x, X) a_n(X)| \leq C \left(E_X |a_n(X)|^s\right)^{\frac{1}{s}}$,

$$\begin{aligned} I_2 \leq & C\, h_+^{-1} \big(E_X |a_n(X)|^s\big)^{\frac{1}{s}} \left(\frac{1}{n}\sum_i |\psi_1(X_i) - \psi_2(X_i)|^Q\right)^{\frac{1}{Q}} \left(\frac{1}{n}\sum_i |b_n(X_i, \epsilon_i|^t\right)^{\frac{1}{t}} \\ \leq & C^*\, h_+^{-1} \|\psi_1 - \psi_2\|_{P_n,Q}.\end{aligned}$$

For K^* given in Lemma 2.1

$$\begin{aligned} & h_+^{-1} E_X \left|\left(K\left(\frac{r_1(X) - r_1(x)}{h}\right) - K\left(\frac{r_2(X) - r_2(x)}{h}\right)\right) f_n(x, X) a_n(X)\right| \\ \leq & h_+^{\frac{1}{s}} \left(h_+^{-1} E_X \left|K^*\left(\frac{r_0(X) - r_0(x)}{h}\right)\right|^{\frac{s}{s-1}}\right)^{\frac{s-1}{s}} \left(E_X |a_n(X)|^s\right)^{\frac{1}{s}} \sum_{k=1}^d \frac{\|r_{1,k} - r_{2,k}\|_\infty}{h_k} \\ \leq & C\, n^{\eta_+/s} \sum_{k=1}^d \frac{\|r_{1,k} - r_{2,k}\|_\infty}{h_k}.\end{aligned}$$

The previous inequality yields an upper-bound on I_3,

$$\begin{aligned} I_3 &\leq C\, n^{\eta_+/s} \sum_{k=1}^{d} \frac{\left\| r_{1,k} - r_{2,k} \right\|_\infty}{h_k} \left(\frac{1}{n} \sum_i |b_n(X_i, \epsilon_i)|^2 \right)^{\frac{1}{2}} \\ &\leq C^*\, n^{\eta_+/s} \frac{1}{d} \sum_{k=1}^{d} \frac{\left\| r_{1,k} - r_{2,k} \right\|_\infty}{h_k}. \end{aligned}$$

This concludes the proof that $\mathcal{C}_n^G$ is a u-cover of $\mathcal{G}$. Next, I bound the empirical diameter of $\mathcal{G}$,

$$\widehat{\mathrm{diam}}_n = \sup_{\theta \in \Theta} \| G(\theta) - G(\theta_0) \|_{P_n}.$$

For every $\theta \in \Theta$ and for K^* given in Lemma 2.1

$$\begin{aligned} & \| G(\theta) - G(\theta_0) \|_{P_n}^2 \\ \leq & C\, n^{-2(\delta-\eta)_{\min}} \frac{1}{n} \sum_i \left\{ h_+^{-1} E_Y \left[K^* \left(\frac{r_0(Y) - r_0(X_i)}{h} \right) |a_n(Y)| \right] |b_n(X_i, \epsilon_i)| \right\}^2 \\ \leq & n^{2(\eta_+/s - (\delta-\eta)_{\min})} \frac{1}{n} \sum_i |b_n(X_i, \epsilon_i)|^2 \leq C_{\mathrm{diam}}^2\, n^{2(\eta_+/s - (\delta-\eta)_{\min})}. \end{aligned}$$

For the constant C from Lemma 5.1 in van de Geer 2000 let

$$\begin{aligned} \beta_n =& 8C\tilde{C}^{\frac{1}{2}} \sum_{j=1}^{d} \left(1 - \gamma_j/2 \right)^{-1} n^{\gamma_j/2(\eta_j + \eta_+/s)} \left(\widehat{\mathrm{diam}}_n \right)^{1-\gamma_j/2} \\ \geq & 8C\tilde{C}^{\frac{1}{2}} \int_0^{\widehat{\mathrm{diam}}_n} H_n^{1/2}(u)\, du \\ \geq & 8C \int_0^{\widehat{\mathrm{diam}}_n} \log^{1/2} \mathcal{N}\left(u, \mathcal{G}, \|\cdot\|_{P_n}\right) du. \end{aligned}$$

The last inequality holds for large n and is due to the fact that $\widehat{\mathrm{diam}}_n < \tilde{u}$ eventually. Lemma 5.1 and equation (5.1) in van de Geer 2000 yield the exponential inequality

$$P\left(A_n \wedge \sup_{\theta \in \Theta} \frac{1}{n} \left| \sum_i \left(G_i(\theta) - G_i(\theta_0) \right) \right| \geq n^{\frac{1}{2}} \beta_n \right) \leq C \exp\left[- \frac{\beta_n^2}{64 C^2 \widehat{\mathrm{diam}}_n^2} \right].$$

To ensure that the right-hand side of the exponential inequality vanishes it suffices to assume that $\xi_j > 0$ for all j which can be done wlog. Let

$$\kappa_j^* = \frac{1}{2} - \frac{\eta_+}{s} + (\delta - \eta)_{\min} - \frac{1}{2}\left(\xi_j + \eta_j \gamma_j \right).$$

The claim about the convergence rate follows by noting that

$$\min_{1 \leq j \leq d} \kappa_j^* \geq \frac{1}{2} - \frac{\eta_+}{s} + (\delta - \eta)_{\min} - \frac{1}{2} \max_{1 \leq j \leq d} \left[\delta_j \gamma_j + \xi_j \right] > \kappa^*$$

and therefore

$$n^{\kappa^*} n^{-\frac{1}{2}} \beta_n \leq C \sum_{j=1}^{d} C_{\mathrm{diam}}^{1-\gamma_j/2} n^{-(\kappa_j^* - \kappa^*)} \to 0.$$

□

Lemma 2.3 *Let Φ and Ψ denote bounded VC-classes. Let $f_n : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote a function satisfying $\limsup_{n\to\infty} \sup_{x_1 \in \mathcal{X}, x_2 \in \mathcal{X}} |f_n(x_1, x_2)| < \infty$ for some $\lambda \in \mathbb{R}$ and let $a_n : \mathcal{X} \to \mathbb{R}$ and $b_n : \mathcal{W} \to \mathbb{R}$ denote functions satisfying*

$$\limsup_{n\to\infty} E\,|a_n(X)|^s < \infty \quad \textit{and} \quad \limsup_{n\to\infty} E\,|b_n(X,\epsilon)|^t < \infty$$

for $s > 2$ and $t > 2$. Define

$$G(\phi,\psi,x,u) = h_+^{-1} E_X \left[K\left(\frac{r_0(X) - r_0(x)}{h} \right) f_n(x,X) a_n(X) \phi(X) \right] b_n(x,u)\psi(x)$$

and let $G_i(\cdot,\cdot) = G(\cdot,\cdot,X_i,\epsilon_i)$ and

$$\kappa^* = \frac{1}{2} - \frac{\eta_+}{s}.$$

Then

$$\sup_{\phi\in\Phi,\psi\in\Psi} \left| \frac{1}{n} \sum_{i=1}^n \big(G_i(\theta) - EG_i(\theta) \big) \right| = O_p\left(n^{-\kappa^*} \sqrt{\log n} \right).$$

Proof Let $\mathcal{G} = \{G(\phi,\psi) : \phi \in \Phi, \psi \in \Psi\}$. First show that there is a positive constant $\tilde{C}$ such that on a set wpa1

$$\log \mathcal{N}\left(u, \mathcal{G}, \|\cdot\|_{P_n}\right) \le \tilde{C}\, H_n(u),$$

with

$$H_n(u) = (\log n) + \log(u^{-1} \vee 1).$$

Let $A_n = \{\|b_n\|_{P_n,s} \le C_A\}$ and choose C_A large enough that $PA_n \to 1$. Now work conditional on A_n. Construct a u-cover of $\mathcal{G}$ in the following way. Take $\tilde{u} = (C^*)^{-1} h_+(u/2)$, where C^* is a constant to be determined later. Now, let $\mathcal{C}_n^\phi$ denote a $\tilde{u}$-cover of Φ with respect to the $\|\cdot\|_{P,\frac{s}{s-1}}$-norm and let $\mathcal{C}_n^\psi$ denote a $\tilde{u}$-cover of Ψ with respect to the $\|\cdot\|_{P_n,Q}$-norm, $Q = \frac{2s}{s-2}$. Since Φ and Ψ are VC, there exist constants A and V such that $\mathcal{C}_n^\phi$ and $\mathcal{C}_n^\psi$ can be chosen to contain less than $A\tilde{u}^{-V}$ functions each (Theorem 2.6.4 in van der Vaart and Wellner 1996). By construction

$$\mathcal{C}_n^G = \left\{ G(\phi,\psi) : \phi \in \mathcal{C}_n^\phi, \psi \in \mathcal{C}_n^\psi \right\}$$

is a cover of $\mathcal{G}$ and can be chosen such that

$$\log \#\mathcal{C}_n^G \le M_1 + M_2 \log n + M_3 \log u^{-1}$$

for constants $M_1, \ldots, M_3$. Next, I show that if C^* is chosen large enough then $\mathcal{C}_n^G$ will be a u-cover of $\mathcal{G}$ with respect to the $\|\cdot\|_{P_n}$-norm. Let $(\phi_1,\psi_1) \in (\Phi,\Psi)$ and let (ϕ_2,ψ_2) denote a nearest grid-point, i.e., $\|\phi_1 - \phi_2\|_{P,\frac{s}{s-1}} \le \tilde{u}$ and $\|\psi_1 - \psi_2\|_{P_n,Q} \le \tilde{u}$. Decompose

$$\begin{aligned} \|G(\phi_1,\psi_1) - G(\phi_2,\psi_2)\|_{P_n} &\le \|G(\phi_1,\psi_1) - G(\phi_2,\psi_1)\|_{P_n} \\ &\quad + \|G(\phi_2,\psi_1) - G(\phi_2,\psi_2)\|_{P_n} \le I_1 + I_2. \end{aligned}$$

To bound the right-hand side note that on A_n

$$\left(\frac{1}{n}\sum_{i=1}^{n}|b_n(X_i,\epsilon_i)|^2\right)^{\frac{1}{2}} \leq \left(\frac{1}{n}\sum_{i=1}^{n}|b_n(X_i,\epsilon_i)|^t\right)^{\frac{1}{t}} \leq C_A.$$

To bound I_1 note that

$$\begin{aligned}&E_X\left|f_n(x,X)a_n(X)\big(\phi_1(X)-\phi_1(X)\big)\right|\\ \leq& C\left(E_X|a_n(X)|^s\right)^{\frac{1}{s}}\left(E_X|\phi_1(X)-\phi_2(X)|^{\frac{s}{s-1}}\right)^{\frac{s-1}{s}}.\end{aligned}$$

Therefore, for C^* chosen large enough

$$\begin{aligned}I_1 \leq& C\,h_+^{-1}\Big(E_X|\phi_1(X)-\phi_2(X)|^{\frac{s}{s-1}}\Big)^{\frac{s-1}{s}}\left(\frac{1}{n}\sum_{i=1}^{n}|b_n(X_i,\epsilon_i)|^2\right)^{\frac{1}{2}}\\ \leq& C^*\,h_+^{-1}\,\|\phi_1-\phi_2\|_{P,\frac{s}{s-1}}\,.\end{aligned}$$

Due to $E_X|f_n(x,X)a_n(X)| \leq C\left(E_X|a_n(X)|^s\right)^{\frac{1}{s}}$,

$$\begin{aligned}I_2 \leq& C\,h_+^{-1}\big(E_X|a_n(X)|^s\big)^{\frac{1}{s}}\left(\frac{1}{n}\sum_i|\psi_1(X_i)-\psi_2(X_i)|^Q\right)^{\frac{1}{Q}}\left(\frac{1}{n}\sum_i|b_n(X_i,\epsilon_i|^t\right)^{\frac{1}{t}}\\ \leq& C^*\,h_+^{-1}\,\|\psi_1-\psi_2\|_{P_n,Q}\,.\end{aligned}$$

This concludes the proof that C^* can be chosen such that $\mathcal{C}_n^G$ is indeed a u-cover of $\mathcal{G}$. Next, I bound the empirical diameter of $\mathcal{G}$,

$$\widehat{\text{diam}}_n = \sup_{\phi\in\Phi,\psi\in\Psi}\|G(\phi,\psi)\|_{P_n}\,.$$

For every $\phi\in\Phi$ and $\psi\in\Psi$ and for K^* given in Lemma 2.1

$$\begin{aligned}&\|G(\phi,\psi)\|_{P_n}^2\\ \leq& C\,\frac{1}{n}\sum_i\left\{h_+^{-1}E_Y\left[K^*\left(\frac{r_0(Y)-r_0(X_i)}{h}\right)|a_n(X)|\right]|b_n(X_i,\epsilon_i)|\right\}^2\\ \leq& n^{2\eta_+/s}\frac{1}{n}\sum_i|b_n(X_i,\epsilon_i|^2 \leq C\,n^{\eta_+/s}.\end{aligned}$$

Note that

$$\int_0^{\widehat{\text{diam}}_n}\sqrt{\log(u^{-1}\vee 1)}\,du \leq \int_0^1\sqrt{\log(u^{-1})}\,du \leq \int_0^1 1+\log u^{-1}\,du \leq 2$$

The second inequality follows from $\sqrt{x} < x$ for $x > 1$ and $\sqrt{x} \leq 1$ for $0 \leq x \leq 1$. Therefore, $\sqrt{x} \leq x \vee 1$. The last inequality holds since $\int_0^x \log(u^{-1})\,du = x\left[x+\log\left(\frac{1}{x}\right)\right]$ for $x > 0$. To bound the covering integral set $\beta_n(u) = n^{\eta_+/s}\sqrt{\log n}$. For C_β large enough

$$\begin{aligned}\int_0^{\widehat{\text{diam}}_n}\sqrt{H_n(u)}\,du \leq& \int_0^{\widehat{\text{diam}}_n}\left(\sqrt{\log n}+\sqrt{\log\left(u^{-1}\vee 1\right)}\right)du\\ \leq& \sqrt{\log n}\;\widehat{\text{diam}}_n + 2 \leq C_\beta\beta_n.\end{aligned}$$

Choose $\tilde{M} > \tilde{C}C_\beta$. Lemma 5.1 in van de Geer 2000 in conjunction with Corollary 3.4 from the same book gives the exponential inequality

$$P\left(A_n \wedge \sup_{\phi\in\Phi,\psi\in\Psi} \frac{1}{n}\left|\sum_i G_i(\phi,\psi)\right| \geq \tilde{M}\, n^{-\frac{1}{2}}\beta_n\right) \leq C\exp\left[-\frac{\tilde{M}^2\beta_n^2}{64C^2\widehat{\operatorname{diam}}_n^2}\right].$$

The conclusion follows by noting that the right-hand side of the exponential inequality vanishes. □

Lemma 2.4 (Nolan and Pollard 1987) *Let $\mathcal{L}$ denote a class of functions on the product space $\mathcal{W}\times\mathcal{W}$ with envelope F satisfying $PL(w,\cdot) = PL(\cdot,w) = 0$ for all $w \in \mathcal{W}$ for all $L \in \mathcal{L}$. For a given sample $W_1,\dots,W_n$ let U_n denote the probability measure that assigns equal mass to each of the $n(n-1)$ ordered pairs of elements from the set $\{W_1,\dots,W_n\}$. There exists a universal constant C such that*

$$\sqrt{n(n-1)}E\sup_{L\in\mathcal{L}}|U_nL| \leq C\,E\big[\theta_n + \tau_n J_n({}^{\theta_n}\!/\!_{\tau_n})\big] \tag{2.2}$$

where

$$J_n(s) = \int_0^s \log\mathcal{N}\left(x,\mathcal{L},\|\cdot\|_{U_{2n}/\tau_n}\right)\,dx,$$
$$\tau_n = \left(U_{2n}F^2\right)^{\frac{1}{2}} \quad \textit{and} \quad \theta_n = \frac{1}{4}\sup_{L\in\mathcal{L}}\left(U_{2n}L^2\right)^{\frac{1}{2}}.$$

Proof The proof proceeds along the steps outlined in Nolan and Pollard 1987 with some small changes first proposed by Sherman 1994. □

Lemma 2.5 *Let Φ and Ψ bounded VC-classes and let ϕ and ψ denote generic element from Φ and Ψ, respectively. For a function $f_n : \mathcal{W}\times\mathcal{W} \to \mathbb{R}$ suppose that $\limsup_{n\to\infty}\sup_{w_1\in\mathcal{W},w_2\in\mathcal{W}}|f_n(w_1,w_2)| < \infty$. For some $\alpha \in \mathbb{N}_0^d$ let*

$$\tilde{K}^\alpha_{h,ij}(r) = K_{h,ij}(r)\left(\frac{r(X_i)-r(X_j)}{h}\right) \qquad \textit{and}$$
$$G_{ij}(\phi,r) = \tilde{K}^\alpha_{h,ij}(r)\,f_n(W_i,W_j)\phi(X_i)\psi(X_j)\epsilon_i.$$

Suppose that

$$\frac{1}{2}\left(1-\eta_+\right) + (\delta-\eta)_{\min} - \max_{1\leq k\leq d}\left[\delta_k\gamma_k + \xi_k\right] > 0$$

and that $\max_{1\leq k\leq d}\gamma_k < 1$. *Let $\Delta_{ij}(\phi,r) = G_{ij}(\phi,r) - G_{ij}(\phi,r_0)$ and define the kernel*

$$\begin{aligned}L_{i,j}(r,\phi,\psi) =& L(r,\phi,\psi,W_i,W_j)\\ =&\Delta_{ij}(r,\phi,\psi) - E_{W_i}\Delta_{ij}(r,\phi,\psi) - E_{W_j}\Delta_{ij}(r,\phi,\psi) + E\Delta_{ij}(r,\phi,\psi).\end{aligned}$$

Then

$$\sup_{r\in\mathcal{R}_n,\phi\in\Phi,\psi\in\Psi}\left|\frac{1}{n(n-1)}\sum_{i\neq j}L_{ij}(r,\phi,\psi)\right| = o_p\left(n^{-\frac{1}{2}}\right).$$

Proof For K^* given in Lemma 2.1 and a constant C_F let

$$\tilde{F}(W_i, W_j) = C_F\, n^{-(\delta-\eta)_{\min}}\left[h_+^{-1} K^* \left(\frac{r_0(X_i) - r_0(X_j)}{h}\right)|\epsilon_i| + |\epsilon_i|^{s/2} + 1\right].$$

A straight-forward application of Lemma 2.1 gives that

$$F(W_i, W_j) = \tilde{F}(W_i, W_j) + E_{W_i}\tilde{F}(W_i, W_j) + E_{W_j}\tilde{F}(W_i, W_j) + +E\,\tilde{F}(W_i, W_j)$$

is an envelope for L, i.e., $|L| \le F$ provided that C_F is chosen large enough. Let $\tau_n = \left(U_{2n}F^2\right)^{\frac{1}{2}}$. Let $\Theta = \mathcal{R}_n \times \Phi \times \Psi$ and let $\theta = (r, \phi, \psi)$ denote a generic element from Θ. Fix any $u > 0$. There exists a constant $\tilde{C}$ that is independent of u and a u-cover of $\mathcal{L} = \{L(\theta) : \theta \in \Theta\}$ that has logarithmic size of less than $\tilde{C}\, H_n(u)$,

$$H_n(u) = \log n + \log u + \sum_{k=1}^{d}\left(n^{\xi_k + \gamma_k(\delta-\eta)_{\min} + \gamma_k\eta_k} u^{-\gamma_k}\right).$$

To construct such a cover let $\tilde{u} = (C^*)^{-1}h_+ n^{-(\delta-\eta)_{\min}}(u/12)$ and for $k = 1, \dots, d$ let $u_{r,k} = (C^*)^{-1}n^{-(\delta-\eta)_{\min}-\eta_k}(u/12)$, where C^* is a constant to be determined later. Work conditionally on a sample of size $2n$ from the distribution of $W = (X, \epsilon)$. Now, take a $\tilde{u}$-cover $\mathcal{C}_n^{\phi}$ of Φ with respect to the $\|\cdot\|_{P_{2n},Q}$-norm, $Q = \frac{2s}{s-2}$, and a a $\tilde{u}$-cover $\mathcal{C}_n^{\phi'}$ of Φ with respect to the $\|\cdot\|_P$-norm. Similarly, take a $\tilde{u}$-cover $\mathcal{C}_n^{\psi}$ of Ψ with respect to the $\|\cdot\|_{P_{2n},Q}$-norm, $Q = \frac{2s}{s-2}$, and a a $\tilde{u}$-cover $\mathcal{C}_n^{\psi'}$ of Ψ with respect to the $\|\cdot\|_P$-norm. Since Φ and Ψ are VC, there are constants A and V such that $\mathcal{C}_n^{\phi}$, $\mathcal{C}_n^{\phi'}$, $\mathcal{C}_n^{\psi}$, and $\mathcal{C}_n^{\psi'}$ can be chosen to contain less than $A\tilde{u}^{-V}$ approximating functions each (Theorem 2.6.4 in van der Vaart and Wellner 1996). Also, for $k = 1, \dots, d$ let $\mathcal{C}_{n,k}^{r}$ denote a minimal $u_{r,k}$-cover of $\mathcal{R}_{k,n}$ with respect to the $\|\cdot\|_\infty$-norm. Let

$$\mathcal{C}_n^L = \Bigg\{\Delta(r, \phi, \psi) - E_{W_i}\Delta(r, \phi', \psi) - E_{W_j}\Delta(r, \phi, \psi') \\ + E\Delta(r, \phi', \psi') : (r, \phi, \phi', \psi, \psi') \in \prod_{k=1}^{d}\mathcal{C}_{n,k}^{r} \times \mathcal{C}_n^{\phi} \times \mathcal{C}_n^{\phi'} \times \mathcal{C}_n^{\psi} \times \mathcal{C}_n^{\psi'}\Bigg\}.$$

It is straightforward to check that there is a $\tilde{C}$ independent of u such that $\log \#\mathcal{C}_n^L \le \tilde{C}H_n(u)$. I will now show that C^* can be chosen so that $\mathcal{C}_n^L$ is a u-cover of $\mathcal{L}$ with respect to the $\|\cdot\|_{U_{2n}/\tau_n}$-norm. Let $L(\theta_1)$, $\theta_1 = (r_1, \phi_1, \psi_1)$, denote an arbitrary point in $\mathcal{L}$ and let

$$\Delta(r_2, \phi_2, \psi_2) - E_{W_i}\Delta(r_2, \phi_2', \psi_2) - E_{W_j}\Delta(r_2, \phi_2, \psi_2') + E\Delta(r_2, \phi_2', \psi_2')$$

denote a nearest grid point in $\mathcal{C}_n^L$, i.e.,

$$\|\phi_1 - \phi_2\|_{P_{2n},Q} \le \tilde{u} \qquad \|\phi_1 - \phi_2'\|_P \le \tilde{u}$$
$$\|\psi_1 - \psi_2\|_{P_{2n},Q} \le \tilde{u} \qquad \|\psi_1 - \psi_2'\|_P \le \tilde{u}$$

and $\|r_{1,j} - r_{2,j}\|_\infty \leq u_{r,j}$ for $j = 1, \ldots, d$. Decompose

$$\begin{aligned}
&\Big|Ł(\theta_1) - \big(\Delta(r_2, \phi_2, \psi_2) - E_{W_i}\Delta(r_2, \phi_2', \psi_2) \\
&\qquad\qquad - E_{W_j}\Delta(r_2, \phi_2, \psi_2') + E\Delta(r_2, \phi_2', \psi_2')\big)\Big|_{U_{2n}/\tau_n} \\
\leq &\|\Delta(r_1, \phi_1, \psi_1) - \Delta(r_2, \phi_1, \psi_1)\|_{U_{2n}/\tau_n} \\
&+ \|\Delta(r_2, \phi_1, \psi_1) - \Delta(r_2, \phi_2, \psi_1)\|_{U_{2n}/\tau_n} \\
&+ \|\Delta(r_2, \phi_2, \psi_1) - \Delta(r_2, \phi_2, \psi_2)\|_{U_{2n}/\tau_n} \\
&+ \|E_{W_i}\Delta(r_1, \phi_1, \psi_1) - E_{W_i}\Delta(r_2, \phi_1, \psi_1)\|_{U_{2n}/\tau_n} \\
&+ \Big\|E_{W_i}\Delta(r_2, \phi_1, \psi_1) - E_{W_i}\Delta(r_2, \phi_2', \psi_1)\Big\|_{U_{2n}/\tau_n} \\
&+ \Big\|E_{W_i}\Delta(r_2, \phi_2', \psi_1) - E_{W_i}\Delta(r_2, \phi_2', \psi_2)\Big\|_{U_{2n}/\tau_n} \\
&+ \Big\|E_{W_j}\Delta(r_1, \phi_1, \psi_1) - E_{W_j}\Delta(r_2, \phi_1, \psi_1)\Big\|_{U_{2n}/\tau_n} \\
&+ \Big\|E_{W_j}\Delta(r_2, \phi_1, \psi_1) - E_{W_j}\Delta(r_2, \phi_2, \psi_1)\Big\|_{U_{2n}/\tau_n} \\
&+ \Big\|E_{W_j}\Delta(r_2, \phi_2, \psi_1) - E_{W_j}\Delta(r_2, \phi_2, \psi_2')\Big\|_{U_{2n}/\tau_n} \\
&+ \|E\,\Delta(r_1, \phi_1, \psi_1) - E\,\Delta(r_2, \phi_1, \psi_1)\|_{U_{2n}/\tau_n} \\
&+ \Big\|E\,\Delta(r_2, \phi_1, \psi_1) - E\,\Delta(r_2, \phi_2', \psi_1)\Big\|_{U_{2n}/\tau_n} \\
&+ \Big\|E\,\Delta(r_2, \phi_2', \psi_1) - E\,\Delta(r_2, \phi_2', \psi_2')\Big\|_{U_{2n}/\tau_n} \\
\leq &I_1 + I_2 + I_3 + I_4 + I_5 + I_6 + I_7 + I_8 + I_9 + I_{10} + I_{11} + I_{12}.
\end{aligned}$$

I will now show that each of the terms $I_1, \ldots, I_{12}$ can be bounded by $u/12$. By Hölder's inequality

$$\begin{aligned}
I_1 \leq &C\,\tau_n^{-1} \sum_{k=1}^{d} \frac{\|r_{1,k} - r_{2,k}\|_\infty}{h_k} \sqrt{\frac{1}{n(n-1)} \sum_{i \neq j} \left(h_+^{-1} K^* \left(\frac{r_0(X_i) - r_0(X_i)}{h} \right) \epsilon_i \right)^2} \\
\leq &C^*\, n^{(\delta-\eta)_{\min}} \frac{1}{d} \sum_{k=1}^{d} \frac{\|r_{1,k} - r_{2,k}\|_\infty}{h_k}.
\end{aligned}$$

The last inequality is due to

$$\tau_n \geq C_F\, n^{-(\delta-\eta)_{\min}} \frac{1}{n(n-1)} \sum_{i \neq j} h_+^{-1} K^* \left(\frac{r_0(X_i) - r_0(X_i)}{h} \right) |\epsilon_i|\,.$$

I_4, I_7 and I_{10} can be bounded in a similar manner.

$$\begin{aligned}
I_2 \leq &C\,\tau_n^{-1} n^{-(\delta-\eta)_{\min}} \sqrt{\frac{1}{n(n-1)} \sum_{i \neq j} \left[h_+^{-1} K^* \left(\frac{r_0(X_i) - r_0(X_i)}{h} \right) |\epsilon_i| \big(\phi_1(X_i) - \phi_2(X_i)\big) \right]^2} \\
\leq &C\,\tau_n^{-1} n^{-(\delta-\eta)_{\min}} h_+^{-1} \left(\frac{1}{n(n-1)} \sum_{i \neq j} |\phi_1(X_i) - \phi_2(X_i)|^Q \right)^{\frac{1}{Q}} \left(\frac{1}{n(n-1)} \sum_{i \neq j} |\epsilon_i|^s \right)^{\frac{1}{s}}
\end{aligned}$$

$$
\begin{aligned}
&\leq C\,\tau_n^{-1} n^{-(\delta-\eta)_{\min}} h_+^{-1} \left\|\phi_1 - \phi_2\right\|_{P_{2n}} \left(\frac{1}{n(n-1)} \sum_{i \neq j} |\epsilon_i|^s + 1 \right)^{\frac{1}{2}} \\
&\leq C^*\, h_+^{-1} \left\|\phi_1 - \phi_2\right\|_{P_{2n}} .
\end{aligned}
$$

Similarly, one can argue that

$$
\begin{aligned}
I_3 \leq& C^*\, h_+^{-1} \left\|\psi_1 - \psi_2\right\|_{P_{2n}} \qquad \text{and} \\
I_8 \leq& C^*\, h_+^{-1} \left\|\phi_1 - \phi_2\right\|_{P_{2n}}
\end{aligned}
$$

By the conditional Cauchy-Schwarz inequality

$$
\left| E_{W_i} \epsilon_i \big(\phi_1(X_i) - \phi_2(X_i)\big) \right| \leq \sqrt{E_{W_i} |\epsilon_i|^2} \sqrt{E_{W_i} \left|\phi_1(X_i) - \phi_2(X_i)\right|^2} .
$$

This gives a bound for I_5

$$
I_5 \leq C^*\, h_+^{-1} \left\|\psi_1 - \psi_2'\right\|_P .
$$

Similar arguments yield

$$
\begin{aligned}
I_{11} \leq& C^*\, h_+^{-1} \left\|\phi_1 - \phi_2'\right\|_P \qquad \text{and} \\
I_{12} \leq& C^*\, h_+^{-1} \left\|\psi_1 - \psi_2'\right\|_P .
\end{aligned}
$$

By the Cauchy-Schwarz and Hölder's inequalities

$$
\begin{aligned}
I_6 \leq& C\,\tau_n^{-1} n^{-(\delta-\eta)_{\min}} h_+^{-1} \left(\frac{1}{n(n-1)} \sum_{i \neq j} \Big[E_{W_i} \epsilon_i (\psi_1(X_j) - \psi_2(X_j)) \Big]^2 \right)^{\frac{1}{2}} \\
\leq& C\,\tau_n^{-1} n^{-(\delta-\eta)_{\min}} h_+^{-1} \left(\frac{1}{n(n-1)} \sum_{i \neq j} \big(\psi_1(X_j) - \psi_2(X_j)\big)^2 E_{W_i} |\epsilon_i|^2 \right)^{\frac{1}{2}} \\
\leq& C\, h_+^{-1} \left(\frac{1}{n(n-1)} \sum_{i \neq j} \big(\psi_1(X_j) - \psi_2(X_j)\big)^Q \right)^{\frac{1}{Q}} \leq C^*\, h_+^{-1} \left\|\psi_1 - \psi_2\right\|_{P_{2n,Q}} .
\end{aligned}
$$

An obvious variation of previous arguments gives

$$
I_9 \leq C^* h_+^{-1} \left\|\psi_1 - \psi_2'\right\|_P .
$$

This concludes the proof that there is a constant $\tilde{C}$ such that

$$
\log \mathcal{N}\left(u, \mathcal{L}, \|\cdot\|_{U_{2n}/\tau_n}\right) \leq \tilde{C}\, H_n(u)
$$

for all u. Now apply Lemma 2.4. Let J and θ_n as defined in Lemma 2.4. Note that it is wlog to assume $\xi_k > 0$ for all $k = 1, \ldots, d$. By assumption there is a constant $a < 1$ such that $0 \leq \gamma_k < a$ for $k = 1, \ldots, d$ and thus eventually

$$
J\left({}^{\theta_n}/_{\tau_n}\right) \leq \tilde{C} \int_0^1 H_n(u)\, du \leq \frac{\tilde{C}}{1-a} \sum_{k=1}^{d} \left(n^{\xi_k + (\delta-\eta)_{\min} + \gamma_k + \eta_k \gamma_k} \right).
$$

The right-hand side of (2.2) is therefore bounded by

$$\frac{2\tilde{C}}{1-a} E \|F\|_{U_{2n}} \sum_{k=1}^{d} \left(n^{\xi_k + (\delta-\eta)_{\min} + \gamma_k + \eta_k \gamma_k} \right).$$

To bound the expectations on the right-hand side write $\tilde{F}_{ij} = \tilde{F}(W_i, W_j)$ and use that $E\|\cdot\|_{U_{2n}} \le \left(E \|\cdot\|_{U_{2n}}^2 \right)^{1/2}$. By the conditional Jensen inequality

$$\begin{aligned} E\|F\|_{U_{2n}}^2 \le & 8 \frac{1}{n(n-1)} \sum_{i \neq j} \left(E\big(\tilde{F}_{ij}\big)^2 + E\big(E_{W_i} \tilde{F}_{ij}\big)^2 + E\big(E_{W_i} \tilde{F}_{ij}\big)^2 + E\big(E\, \tilde{F}_{ij}\big)^2 \right) \\ \le & 8 \frac{1}{n(n-1)} \sum_{i \neq j} \left(E\big(\tilde{F}_{ij}\big)^2 + E\, E_{W_i} \big(\tilde{F}_{ij}\big)^2 + E\, E_{W_i} \big(\tilde{F}_{ij}\big)^2 + E\big(\tilde{F}_{ij}\big)^2 \right) \\ \le & \frac{1}{n(n-1)} \sum_{i \neq j} 32 E\big(\tilde{F}_{ij}\big)^2. \end{aligned}$$

There is a constant C such that

$$\begin{aligned} E\big(\tilde{F}_{ij}\big)^2 \le & 4\, C_F^2\, n^{-2(\delta-\eta)_{\min}} \left(E \left| h_+^{-1} K^* \left(\frac{r_0(X_i) - r_0(X_j)}{h} \right) \right|^2 + E\, |\epsilon_i|^s + 1 \right) \\ \le & C\, n^{-2(\delta-\eta)_{\min}} h_+^{-1}. \end{aligned}$$

The last inequality follows since

$$h_+^{-1} E_X \left[K^* \left(\frac{r_0(x) - r_0(X)}{h} \right) \right]^2 \le \int |K^*(v)|^2 f_{r_0(X)}(r_0(x) - vh)\, dv \le C$$

for a constant C that does not depend on x. Collecting the bounds from the previous displays gives

$$E\|F\|_{U_{2n}} \le C\, n^{\frac{1}{2}\eta_+ - (\delta-\eta)_{\min}}.$$

Employing inequality (2.2) of Lemma 2.4 yields

$$\sqrt{n} E \sup_{L \in \mathcal{L}} |U_n L| \le C\ (n-1)^{-\frac{1}{2}}\, n^{\frac{1}{2}\eta_+ - (\delta-\eta)_{\min}} \sum_{k=1}^{d} \left(n^{\xi_k + \gamma_k (\delta-\eta)_{\min} + \gamma_k \eta_k} \right).$$

The right-hand side vanishes if $\min_{1 \le k \le d} \kappa_k > 0$, where

$$\kappa_k = \frac{1}{2}\big(1 - \eta_{\min}\big) + (1-\gamma_k)(\delta-\eta)_{\min} - \xi_k - \eta_k \gamma_k \quad \text{for } k = 1, \ldots, d.$$

The desired rate follows from the inequality

$$\kappa_k \ge \frac{1}{2}\big(1 - \eta_{\min}\big) + (\delta-\eta)_{\min} - \max_{1 \le l \le d} \big[\delta_l \gamma_l + \xi_l\big] \quad \text{for } k = 1, \ldots, d.$$

Finally, Markov's inequality converts convergence in mean into convergence in probability, concluding the proof. □

Lemma 2.6 *Let Φ and Ψ denote bounded VC-classes and let ϕ and ψ denote generic elements from Φ and Ψ, respectively. Suppose there are functions $f_n : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ and $b_n : \mathcal{W} \to \mathbb{R}$ satisfying*

$$\limsup_{n\to\infty} \sup_{w_1 \in \mathcal{W}, w_2 \in \mathcal{W}} |f_n(w_1, w_2)| < \infty \quad \textit{and} \quad \limsup_{n\to\infty} E\,|b_n(X, \epsilon)|^2 < \infty.$$

For some $\alpha \in \mathbb{N}_0^d$ let

$$\begin{aligned} \tilde{K}^{\alpha}_{h,ij}(r) &= K_{h,ij}(r) \left(\frac{r(X_j) - r(X_i)}{h} \right) \qquad \textit{and} \\ G_{ij}(\phi, \psi) &= \tilde{K}^{\alpha}_{h,ij}(r_0)\, \phi(X_i)\psi(X_j) f_n(W_i, W_j) b_n(X_i, \epsilon_i). \end{aligned}$$

Moreover, suppose that $\max_{1 \le k \le d} \gamma_k < 1$ and that the density $f_{r_0(X)}$ is bounded. Define the kernel

$$L_{i,j}(\phi,\psi) = L(\phi,\psi,X_i,X_j) = G_{ij}(\phi,\psi) - E_{W_i} G_{ij}(\phi,\psi) - E_{W_j} G_{ij}(\phi,\psi) + E G_{ij}(\phi,\psi)$$

and fix a constant κ such that

$$\kappa < \frac{1}{2}(1 - \eta_+)$$

Then

$$\sqrt{n} \sup_{\phi\in\Phi, \psi\in\Psi} \left| \frac{1}{n(n-1)} \sum_{i \neq j} L_{ij}(\phi,\psi) \right| = o_p\left(n^{-\kappa}\right).$$

Proof It is convenient to consider the scaled process $\tilde{G} = n^{-\frac{1}{2}\eta_+} G$ and the corresponding scaled kernel $\tilde{L} = n^{-\frac{1}{2}\eta_+} L$. As $\mathcal{L} = \{L(\phi,\psi) : \phi \in \Phi, \psi \in \Psi\}$ is VC and hence Euclidean, the Main Corollary in Sherman 1994 is applicable. It suffices to find an envelope $F = F(W_i, W_j)$ such that $|\tilde{G}| \le F$ and $E\,F^2 < \infty$. For a constant C_F and K^* given by Lemma 2.1 let

$$\tilde{F}_{ij} = C_F\, h_+^{-\frac{1}{2}} K^* \left(\frac{r_0(X_i) - r_0(X_j)}{h} \right) |b_n(X_i, \epsilon_i)|.$$

Provided that C_F is chosen large enough, the desired envelope is given by

$$F(W_i, W_j) = \tilde{F}_{ij} + E_{W_i}\tilde{F}_{ij} + E_{W_j}\tilde{F}_{ij} + E\,\tilde{F}_{ij}.$$

To verify the integrability condition write

$$\begin{aligned} E(F(W_i,W_j)^2 \le& 8\,E\Big(\tilde{F}_{ij}^2 + (E_{W_i}\tilde{F}_{ij})^2 + (E_{W_j}\tilde{F}_{ij})^2 + (E\,\tilde{F}_{ij})^2\Big) \le 32\,E\tilde{F}_{ij}^2 \\ \le& 32\,C_F^2\,E\left\{ E_{W_j} h_+^{-1} \left[K^* \left(\frac{r_0(X_i) - r_0(X_j)}{h} \right) \right]^2 |b_n(X_i,\epsilon_i)|^2 \right\} \\ \le& 32\,C_F^2\,E\,|b_n(X_i,\epsilon_i)|^2 < \infty. \end{aligned}$$

The second inequality is due to the conditional version of Jensen's inequality. The third inequality exploits the fact that the bounded density assumption implies boundedness of $E_{W_j} h_+^{-1}\left[K^*\left(\frac{r_0(X_i)-r_0(X_j)}{h}\right)\right]^2$. The Main Corollary in Sherman 1994 gives that the U-process with the scaled kernel is $O_p(n^{-1})$. The convergence rate for the unscaled process is now obvious. □

Lemma 2.7 *Suppose that $m : z \mapsto E[Y \mid r_0(X) = z]$ has $q+1$ bounded derivatives and that the density function $f_{r_0(X)}$ is bounded. Let*

$$G_{ij}(x) = K_{h,ij}(r_0)\,\delta_x(X_i) \sum_{1\le|\alpha|\le q} \frac{h^\alpha}{\alpha!}\partial^\alpha m(r_0(X_j))\left(\frac{r_0(X_i)-r_0(X_j)}{h}\right)^\alpha .$$

Then

$$\sqrt{n}\sup_{x\in\mathbb{R}^{d_x}}\left|\frac{1}{n(n-1)}\sum_{i\neq j}\left(G_{ij}(x) - E\,G_{ij}(x)\right)\right| = O_p\left(n^{-\eta_{\min}}\right).$$

Proof Fix $\alpha \in \mathbb{N}_0^d$ such that $1 \le |\alpha| \le q$. For K^* as given in Lemma 2.1 let

$$\tilde{G}_{ij}^\alpha(x) = n^{\eta_{\min}} K_{h,ij}(r_0)\,\delta_x(X_i)\frac{h^\alpha}{\alpha!}\,\partial^\alpha m(r_0(X_j))\mathbf{1}_{\{K^*_{h,ij}(r_0)>0\}}\left(\frac{r_0(X_i)-r_0(X_j)}{h}\right)^\alpha .$$

Applying Lemma 2.6 with the bounded f_n-function

$$f_n(W_i, W_j) = \frac{h^\alpha}{\alpha!}\,\partial^\alpha m(r_0(X_j))\mathbf{1}_{\{K^*_{h,ij}(r_0)>0\}}\left(\frac{r_0(X_i)-r_0(X_j)}{h}\right)^\alpha$$

yields

$$\sqrt{n}\sup_{x\in\mathbb{R}^{d_x}}\frac{1}{n(n-1)}\left|\sum_{i\neq j}\left(\tilde{G}_{ij}^\alpha(x) - E_{W_i}\tilde{G}_{ij}^\alpha(x) \right.\right.$$
$$\left.\left. - E_{W_j}\tilde{G}_{ij}^\alpha(x) + E\,\tilde{G}_{ij}^\alpha(x)\right)\right| = O_p\left(n^{-\lambda}\right)$$

for all $\lambda < \frac{1}{2}(1-\eta_+)$. Next, establish that

$$\sup_{x\in\mathbb{R}^{d_x}}\left|\frac{1}{n(n-1)}\sum_{i\neq j}\left(E_{W_i}\tilde{G}_{ij}^\alpha(x) - E\,\tilde{G}_{ij}^\alpha(x)\right)\right| = O_p\left(n^{-\frac{1}{2}}\right)$$
$$\sup_{x\in\mathbb{R}^{d_x}}\left|\frac{1}{n(n-1)}\sum_{i\neq j}\left(E_{W_j}\tilde{G}_{ij}^\alpha(x) - E\,\tilde{G}_{ij}^\alpha(x)\right)\right| = O_p\left(n^{-\frac{1}{2}}\right).$$

This follows by applying Lemma 2.3 with the bounded f_n-function from above. The conclusion follows by noting that

$$G_{ij} = n^{-\eta_{\min}}\sum_{1\le|\alpha|\le q}\tilde{G}_{ij}^\alpha .$$

□

Appendix 2.D Local polynomial estimator

Let Y denote a real-valued outcome and let X denote a covariate vector that takes values in $\mathbb{R}^{d_x}$. Suppose that a sample $(Y_i, X_i)_{1\le i\le n}$ from (Y, X) is available. Let $L : \mathbb{R}^{d_x} \to \mathbb{R}$ denote a multi-variate kernel function. Also, let $g = (g_1, \dots, g_{d_x})$ denote a bandwidth sequence, write $g_+ = g_1 \cdots g_{d_x}$ and suppose that $g_j \asymp n^{-\eta_j^*}$. Write $L_g(x) = g_+^{-1} L(x/g)$. The p-th order local polynomial estimator of $r_0(x)$ is given by $\hat{r}(x) = \hat{b}_{(0,\dots,0)'}(x)$ where

$$(\hat{b}_\alpha)_{0\le|\alpha|\le p} \in \arg\min_{b_\alpha(x), 0\le|\alpha|\le p} \frac{1}{n}\sum_i \left[Y_i - \sum_{0\le|\alpha|\le p} g^\alpha b_\alpha(x) \left(\frac{X_i - x}{g}\right)^\alpha \right]^2 L_g(X_i - x).$$

We need a way to order collections of functions indexed by a multi-index α. For a given $j = 1, \dots, p$ there are $N_j = \binom{j+d_x-1}{d_x-1}$ different $\alpha \in \mathbb{N}_0^{d_x}$ such that $|\alpha| = j$. Order these α's lexicographically (with highest priority being given to the last position). Let g_j^{-1} denote the bijection that maps an α with $|\alpha| = j$ into its rank according to the lexicographic ordering. Now, order the α's as follows

$$\alpha_{g_0(1)}, \alpha_{g_1(1)}, \dots, \alpha_{g_1(N_1)}, \dots, \alpha_{g_p(1)}, \dots, \alpha_{g_p(N_p)}.$$

Let $N = N_1 + \cdots + N_p$. Let $\gamma : \{1, \dots, N\} \to \mathbb{N}_0^{d_x}$ denote the function that maps positions in the ordering to the corresponding multi-index

$$\gamma : k \mapsto g_{\max\{j : N_0 + \cdots + N_j \ge k\}}\big(k + 1 - \max\{j : N_0 + \cdots + N_j \ge k\}\big).$$

Also, for any d_x-vector x let $\Gamma(x)$ denote the vector of polynomials given by

$$\big(\Gamma(x)\big)_j = x^{\gamma(j)}$$

and let $b_{0,n}$ and $\hat{b}_n$ denote the N-vectors given by

$$\big(b_{0,n}\big)_j = \frac{g^{\gamma(j)}}{\gamma(j)!}\partial^{\gamma(j)} r \quad \text{and} \quad \big(\hat{b}_n\big)_j = \hat{b}_{\gamma(j)}.$$

The estimated coefficients can be written as

$$\begin{aligned}
&\hat{b}_n(x) - b_{0,n}(x) \\
=&\mathbf{S}_n^{-1}(x)\frac{1}{n}\sum_i \epsilon_i \Gamma\left(\frac{X_i - x}{g}\right) L_g(X_i - x) \\
&+ \mathbf{S}_n^{-1}(x)\frac{1}{n}\sum_i \sum_{|\beta|=p+1} \frac{g^\beta}{\beta!}\partial^\beta r(x)\Gamma\left(\frac{X_i - x}{g}\right)\left(\frac{X_i - x}{g}\right)^\beta L_g(X_i - x) + \mathbf{S}_n^{-1}(x) R_n(x)
\end{aligned}$$

where

$$\begin{aligned} R_n(x) =&(p+1)\sum_{|\beta|=p+1}\frac{g^\beta}{\beta!}\,\frac{1}{n}\sum_i\left(\frac{X_i-x}{g}\right)^\beta\Gamma\left(\frac{X_i-x}{g}\right)L_g(X_i-x),\\ &\qquad\qquad\int_0^1\left[\partial^\beta r(x+\lambda(X_i-x))-\partial^\beta r(x)\right](1-\lambda)^p\,d\lambda\\ \mathbf{S}_n(x) =&\frac{1}{n}\sum_i\Gamma\left(\frac{X_i-x}{g}\right)\left\{\Gamma\left(\frac{X_i-x}{g}\right)\right\}'L_g(X_i-x). \end{aligned}$$

Let $\mathcal{D}$ denote a compact subset of the support of the random variable X and let

$$Q_n(x)=\frac{1}{n}\sum_i\sum_{|\beta|=p+1}\frac{g^\beta}{\beta!}\partial^\beta r(x)\,\Gamma\left(\frac{X_i-x}{g}\right)\left(\frac{X_i-x}{g}\right)^\beta L_g(X_i-x).$$

Under some minor regularity conditions, it is well known (Masry 1996) that

$$\begin{aligned} \sup_{x\in\mathcal{D}}\left|\mathbf{S}_n(x)-E\,\mathbf{S}_n(x)\right| =&O_p\left(\sqrt{\frac{\log n}{ng_+}}\right)\\ \sup_{x\in\mathcal{D}}\left|[E\,\mathbf{S}_n(x)]^{-1}\right| =&O\left(1\right)\\ \sup_{x\in\mathcal{D}}\left|\frac{1}{n}\sum_i\epsilon_i\,\Gamma\left(\frac{X_i-x}{g}\right)L_g(X_i-x)\right| =&O_p\left(\sqrt{\frac{\log n}{ng_+}}\right)\\ \sup_{x\in\mathcal{D}}\left|Q_n(x)-E\,Q_n(x)\right| =&O_p\left(n^{-(p+1)\eta^*_{\min}}\sqrt{\frac{\log n}{ng_+}}\right)\\ \sup_{x\in\mathcal{D}}\left|E\,Q_n(x)\right| =&O\big(n^{-(p+1)\eta^*_{\min}}\big)\\ \sup_{x\in\mathcal{D}}\left|R_n(x)-E\,R_n(x)\right| =&O_p\left(n^{-(p+1)\eta^*_{\min}}\sqrt{\frac{\log n}{ng_+}}\right)\\ \sup_{x\in\mathcal{D}}\left|E\,R_n(x)\right| =&o\big(n^{-(p+1)\eta^*_{\min}}\big). \end{aligned}$$

This gives rise to two useful stochastic expansions. Let $B_n(x)=[E\mathbf{S}_n(x)]^{-1}EQ_n(x)$ and $B^*_n(x)=B_n(x)+[E\mathbf{S}_n(x)]^{-1}ER_n(x)$. Then,

$$\begin{aligned} \sup_{x\in\mathcal{D}}&\left|\hat{b}_n(x)-\left(b_{0,n}(x)+[E\mathbf{S}_n(x)]^{-1}\frac{1}{n}\sum_i\epsilon_i\,\Gamma\left(\frac{X_i-x}{g}\right)L_g(X_i-x)+B_n(x)\right)\right|\\ &\qquad=O_p\left(\frac{\log n}{ng_+}+n^{-(p+1)\eta^*_{\min}}\sqrt{\frac{\log n}{ng_+}}\right)+o\big(n^{-(p+1)\eta^*_{\min}}\big). \end{aligned}$$

In addition we have

$$\sup_{x\in\mathcal{D}}\left|\hat{b}_n(x)-\left(b_{0,n}(x)+[E\mathbf{S}_n(x)]^{-1}\frac{1}{n}\sum_i\epsilon_i\,\Gamma\left(\frac{X_i-x}{g}\right)L_g(X_i-x)+B_n^*(x)\right)\right|$$
$$=O_p\left(\frac{\log n}{ng_+}+n^{-(p+1)\eta^*_{\min}}\sqrt{\frac{\log n}{ng_+}}\right).$$

The second expansion is more suited to our purposes. Next, I focus on

$$\tilde{r}(x)=r_0(x)+e_1'[E\mathbf{S}_n(x)]^{-1}\frac{1}{n}\sum_i\epsilon_i\,\Gamma\left(\frac{X_i-x}{g}\right)L_g(X_i-x)+e_1'B_n^*(x).$$

The $\hat{r}-\tilde{r}$ part is typically so small that it can be dealt with by direct methods. It suffices to bound the covering number of a class containing the random function

$$\varphi_n(x)=\frac{1}{n}\sum_i\epsilon_i\,\Gamma\left(\frac{X_i-x}{g}\right)L_g(X_i-x).$$

This is because adding a function or multiplying by a bounded function does not change the order of the covering number. To ease the notation suppose that $\eta_1^*=\cdots=\eta_{d_x}^*=\eta^*$. Suppose that L has K bounded derivatives and that the derivatives of order K satisfy a Hölder condition with coefficient θ, i.e., for $|\alpha|=K$ and $0<\theta\le 1$

$$\left|\partial^\alpha L(y)-\partial^\alpha L(y')\right|\le C\left\|y-y'\right\|^\theta.$$

Also, let $\mathcal{D}\subset\mathbb{R}^{d_x}$ denote a bounded and convex set. Note that for any $|\alpha|\ge 1$ and on the set $L_g(X_i-x)>0$

$$\left|\partial^\alpha\Gamma\left(\frac{X_i-x}{g}\right)\right|\le C\,n^{|\alpha|\eta^*}\qquad\text{and}$$
$$\left|\partial^\alpha g_+^{-1}L\left(\frac{X_i-x}{g}\right)\right|\le C\,n^{|\alpha|\eta^*}g_+^{-1}\left|\partial^\alpha L\left(\frac{X_i-x}{g}\right)\right|.$$

Under some regularity conditions (e.g. bounded support of L) this implies

$$n^{-|\alpha|\eta^*}\frac{1}{n}\sum_i\epsilon_i\,\partial^\alpha\left\{\Gamma\left(\frac{X_i-x}{g}\right)L_g(X_i-x)\right\}=O_p\left(\sqrt{\frac{\log n}{ng_+^{-1}}}\right).$$

Therefore, wpa1 φ_n is contained in the class of functions whose $|\alpha|$-order derivatives are bounded by $M_{|\alpha|}=n^{m_{|\alpha|}}$ with

$$m_{|\alpha|}>|\alpha|\,\eta^*+\frac{1}{2}d_x\eta^*-\frac{1}{2}.$$

Similarly, it can be shown that for a universal constant C and φ_n in a set of probability approaching one for all $|\alpha|=K$ and any x, x'

$$\left|\frac{\partial^\alpha\varphi_n(x)-\partial^\alpha\varphi_n(x')}{\left\|x-x'\right\|^\theta}\right|\le C\,M_{K+\theta}.$$

Theorem 2.7.1 of van der Vaart and Wellner 1996 implies wpa1 $\varphi_n \restriction_{\mathcal{D}}$ lies in a set Φ satisfying

$$\log \mathcal{N}(\epsilon, \Phi, \|\cdot\|_\infty) \leq C \left(\frac{M}{\epsilon}\right)^{\frac{d}{K+\theta}}$$

for a constant C depending only on K and $\mathcal{D}$. Therefore, wpa1 the random function $\tilde{r} \restriction_{\mathcal{D}}$ satisfies the complexity condition in Assumption 2.2 with $\gamma = \frac{d_x}{K+\theta}$ and

$$\xi > \left((K+\theta)\eta^* - \frac{1}{2}\left[1 - d_x \eta^*\right]\right)\gamma.$$

Suppose that η and η^* are chosen to be the MISE-optimal rates so that $\eta = (d+2(q+1))^{-1}$ (if $q = q_1 = q_2$ in Assumption 2.1) and $\eta^* = (d_x + 2(p+1))^{-1}$, respectively. Then,

$$\delta = \frac{p+1}{d_x + 2(p+1)} \qquad \text{and} \qquad \delta - \eta = \frac{p+1}{d_x + 2(p+1)} - \frac{1}{d + 2(q+1)}.$$

References

Chen, Song Xi and Ingrid Van Keilegom (2009). “A goodness-of-fit test for parametric and semi-parametric models in multiresponse regression”. In: *Bernoulli* 15.4, pp. 955–976.

Delgado, Miguel A and Wenceslao González Manteiga (2001). “Significance testing in nonparametric regression based on the bootstrap”. In: *Annals of Statistics*, pp. 1469–1507.

Dzemski, Andreas and Florian Sarnetzki (2014). “Overidentification test in a nonparametric treatment model with unobserved heterogeneity”. Working Paper.

Escanciano, Juan Carlos, David Jacho-Chávez, and Arthur Lewbel (2014). “Uniform convergence of weighted sums of non and semiparametric residuals for estimation and testing”. In: *Journal of Econometrics* 178, pp. 426–443.

Escanciano, Juan Carlos and Kyungchul Song (2010). “Testing single-index restrictions with a focus on average derivatives”. In: *Journal of Econometrics* 156.2, pp. 377–391.

Hansen, Bruce (2008). “Uniform convergence rates for kernel estimation with dependent data”. In: *Econometric Theory* 24.03, pp. 726–748.

Härdle, Wolfgang and James Marron (1985). “Optimal bandwidth selection in nonparametric regression function estimation”. In: *The Annals of Statistics*, pp. 1465–1481.

Ichimura, Hidehiko (1993). “Semiparametric least squares (SLS) and weighted SLS estimation of single-index models”. In: *Journal of Econometrics* 58.1, pp. 71–120.

Jones, Chris, James Marron, and Simon Sheather (1996). “A brief survey of bandwidth selection for density estimation”. In: *Journal of the American Statistical Association* 91.433, pp. 401–407.

Klein, Roger W and Richard H Spady (1993). “An efficient semiparametric estimator for binary response models”. In: *Econometrica*, pp. 387–421.

Maistre, Samuel and Valentin Patilea (2014). “Nonparametric model checks for single-index assumptions”. Working Paper.

Mammen, Enno, Christoph Rothe, and Melanie Schienle (2012). “Nonparametric regression with nonparametrically generated covariates”. In: *The Annals of Statistics* 40.2, pp. 1132–1170.

— (2015). “Semiparametric estimation with generated covariates”. In: *Econometric Theory.*

Masry, Elias (1996). “Multivariate local polynomial regression for time series: uniform strong consistency and rates”. In: *Journal of Time Series Analysis* 17.6, pp. 571–599.

Nolan, Deborah and David Pollard (1987). “U-processes: Rates of Convergence”. In: *The Annals of Statistics*, pp. 780–799.

Pollard, David (1984). *Convergence of stochastic processes.* Springer.

Sherman, Robert P (1994). “Maximal inequalities for degenerate U-processes with applications to optimization estimators”. In: *The Annals of Statistics*, pp. 439–459.

Stute, Winfried and Li-Xing Zhu (2005). “Nonparametric checks for single-index models”. In: *Annals of Statistics*, pp. 1048–1083.

van de Geer, Sara (2000). *Empirical Processes in M-estimation.* Vol. 6. Cambridge University Press.

van der Vaart, Aad and Jon Wellner (1996). *Weak Convergence and Empirical Processes.* Springer.

Vytlacil, Edward (2002). “Independence, monotonicity, and latent index models: An equivalence result”. In: *Econometrica* 70.1, pp. 331–341.

Xia, Yingcun et al. (2004). “A goodness-of-fit test for single-index models”. In: *Statistica Sinica* 14.1, pp. 1–28.

CHAPTER 3

An empirical model of dyadic link formation in a network with unobserved heterogeneity

3.1 Introduction

Economic agents concentrate a substantial amount of their activities within their networks of interpersonal relationships. These interpersonal relationships play a prominent role when centralized institutions such as markets are missing or unable to provide certain goods or services. Studying them provides valuable insights into many relevant economic problems, such as information dissemination in small communities (Banerjee et al. 2013) and informal insurance (Fafchamps and Lund 2003). Interpersonal relationships can be formalized as directed links between agents. The collection of all links is called the network. Given their vital role in many policy-relevant problems, it is important to understand how networks are formed. Consequently, econometricians have endeavored to estimate models of formation of informal insurance networks in villages (Fafchamps and Gubert 2007; Leung 2014) or friendship networks in high-schools (Mele 2013).

This paper contributes to the literature by offering a new empirical model of network formation. Similar to the classical approach by Holland and Leinhardt 1981, link formation is modelled as a binary choice. An agent establishes a directed link to another agent if, considering the joint attributes of the pair, the link surplus is deemed large enough. Conditional on agent attributes, links are formed independently of each other. This is the defining property of the class of so-called dyadic models. Though frequently applied in practice (Mayer and Puller 2008; Fafchamps and Gubert 2007), little work has been done to understand their theoretical properties (Graham 2014).

The main innovation of my model is that it employs a fixed effects approach to account for relevant attributes that are not observable to the econometrician. Adding fixed effects substantially complicates inference by introducing a so-called incidental parameter problem (Neyman and Scott 1948). As a result, confidence intervals computed from maximum likelihood estimators are not centered at the true parameter values. I investigate this problem formally in an asymptotic framework that sends the number of agents to infinity. For the estimands considered in this paper I provide explicit correction formulas that can be used to center the respective maximum likelihood estimator at the true parameter value.

Most available alternatives to my approach capture unobserved heterogeneity by random effects (Hoff 2005; Duijn, Snijders, and Zijlstra 2004; Krivitsky et al. 2009). A random effects assumption imposes a very simple structure on unobserved heterogeneity and it does not admit correlations between observed and unobserved agent characteristics. Fixed effects dispose of such restrictions and allow for very general unobserved heterogeneity.

My model can capture two features that are frequently observed in real-world networks. *Homophily* refers to the tendency of agents to initiate ties to agents who share similar observed characteristics (McPherson, Smith-Lovin, and Cook 2001). This can be interpreted as a distaste for social distance and is related to the concept of assortative matching in other areas of economics (Becker 1973). *Degree heterogeneity* refers to the fact that agents can exhibit vast differences in the number of in-bound or out-bound links. In my model, degree heterogeneity is driven by homophily as well as by differences in the ability of agents to initiate ties (productivity) and to attract links from other agents (popularity). Due to the fixed effects approach, determinants of productivity and popularity need not be observed, allowing observationally equivalent agents to exhibit diverse linking strategies. This contributes crucially to the ability of the model to disentangle homophily from unobserved sources of degree heterogeneity (Graham 2014).

A researcher might be interested in the linking model for two reasons. For some research questions the bilateral linking model itself is of interest. For example, it might be interesting to investigate whether homophily preferences discriminate against minorities. In other cases, the researcher wants to learn about the behavior of the stochastic network induced by the sum of all bilateral linking decisions. For example, the level of segregation in the network determines how fast information spreads or how susceptible a community is to outbreaks of sexual diseases (Bearman, Moody, and Stovel 2004).

To my knowledge, I am the first to formally discuss inference on local or global structure of the network in the context of a dyadic network model. On the population level, it is straightforward to calculate various features of the network from a known bilateral linking model. This simplicity does not extend to estimation. In the present paper, this is illustrated by a detailed discussion of a measure of transitivity. The level of transitivity observed in a network is driven by agent productivity and popularity, i.e., the agent-level heterogeneity captured by the fixed effects. Expected transitivity is therefore a function of the fixed effects. Estimates of the fixed effects are provided as a by-product of my estimation procedure allowing for a simple plug-in estimator. This highlights an advantage of my method over alternative approaches in the literature on non-linear models that condition out the fixed effects (Andersen 1970; Charbonneau 2014). However, the plug-in estimator is affected by an incidental parameter problem, rendering standard inference invalid. For the transitivity measure, I propose a procedure that overcomes this limitation by adjusting for asymptotic bias and by estimating robust standard errors. The general approach can be extended to other network features of interest, such as average degree or various clustering coefficients.

Comparing predicted network features to their observed counterparts can serve as a test of model specification. This paper considers such a test based on predicted transitivity. The test can be interpreted as looking in the direction of alternatives in which transitive relationships have explanatory power. The suggested procedure expands on the idea of the τ^2-test in Holland and Leinhardt 1978 by allowing for an estimated reference distribution. The estimation of model parameters induces an incidental parameter problem for the test statistic. My testing procedure accounts for the presence of incidental parameters and produces asymptotically valid critical values. For existing transitivity tests (Holland and Leinhardt 1981; Karlberg 1997; Karlberg 1999) there are no formal results regarding their asymptotic distribution. This paper provides for the first time a large sample theory for a transitivity test for networks.

The finite-sample properties of my methods are investigated in simulations. In my simulation design, the correction formulas offer considerable improvements. The empirical coverage of confidence intervals constructed from uncorrected maximum likelihood estimates is up to sixty percentage points below the nominal level. Applying the correction formulas substantially increases the precision of the estimators and eliminates bias almost completely. This results in an improved normal approximation that produces confidence intervals that hold their nominal coverage level.

Identification in my model is achieved by an exogeneity assumption. Agents evaluate each potential link in isolation of the rest of the network. In particular, there are no *network externalities*. This means that linking decisions are independent of endogenous network structure. This is plausible if agents do not care about links between other agents or if the network is imperfectly observable. The exogeneity assumption is refutable by the model specification test developed in this paper.

The literature on network formation offers some models that allow for network externalities. These models do not, however, admit general unobserved heterogeneity. For some facets of network structure, such as transitivity, network externalities and unobserved heterogeneity offer competing explanations. To estimate a game of network formation under asymmetric information, Leung 2014 provides a model in the spirit of Aguirregabiria and Mira 2007. His approach can account for network externalities but it requires observationally identical agents to play identical strategies. My model does not constrain heterogeneity in this way. In applied research, exponential random graph models (Wasserman and Pattison 1996; Snijders et al. 2006) are a popular way to endogenize local network structure. Their micro-foundation (Mele 2013) does not permit unobserved heterogeneity, they can be computationally intractable (Bhamidi, Bresler, and Sly n.d.) and frequentist properties of estimators based on these models are largely unknown (Chandrasekhar and Jackson 2014). My model does not impose such restrictions.

Conditional on observed and unobserved agent characteristics, the stochastic network induced by my dyadic linking model is an Erdős-Rényi graph (Erdős and Rényi 1960). In real-world networks, unconditional or conditional-on-observables Erdős-Rényi models often understate the level of transitivity (Davis 1970; Watts and Strogatz 1998; Apicella et al. 2012). This is commonly attributed to the presence of network externalities and taken to indicate that agents derive utility from transitive closure. In the context of a stylized example I offer an alternative explanation for the puzzle by showing that the omission of latent popularity effects will lead to a downward bias of predicted transitivity.

The relevance of unobserved heterogeneity in a real-world network is investigated in an empirical application. In the application, the methods developed in this paper are applied to data on favor networks in Indian villages. The favor networks are constructed from the survey data of Jackson, Rodriguez-Barraquer, and Tan 2012 and Banerjee et al. 2013. A directed link from agent i to agent j exists if i nominates j as someone she would ask for help if she needed to borrow household staples or money. From an economic perspective these relationships are interesting because they can serve as a partial insurance device. Predictions for transitivity from the model with fixed effects are compared to predictions from a simple linking model in which linking decisions are based solely on observed characteristics. The model with fixed effects predicts a much higher level of transitivity than the simple model. Notably, the level of transitivity observed in the sampled networks exceeds the predictions from the simple model by a significant amount. In contrast, under the model with fixed effects, the transitivity test does not detect excess transitivity. These results suggest that unobserved agent effects may affect the evolution of the favor networks in a substantial way. In particular, controlling for unobserved effects is essential for replicating the observed level of transitivity. This can be achieved by using the methods developed in this paper.

In parallel research, Graham 2014 develops a dyadic network model with fixed effects. My research differs from his contribution in two ways. First, I consider directed links, whereas Graham 2014's model assumes an undirected network. The choice of model is dictated by the nature of the available data. Without data on the direction of links, productivity and popularity effects can not be distinguished. In my application, there is no monotone relationship between the two effects, suggesting a complex heterogeneity pattern that would not be captured well by the kind of one-dimensional heterogeneity that an undirected model is limited to. Secondly, Graham 2014 focuses on estimation of the homophily component of link surplus, whereas I also discuss estimation and testing of local structure.

From a technical perspective, dyadic network models are closely related to long-T panel models. Consequently, this research ties in with the recent literature on incidental parameters in non-linear panel models (Hahn and Kuersteiner 2011; Hahn and Newey 2004). In particular, some of the theoretical insights presented in this paper build on results for maximum likelihood models with incidental parameters in Fernández-Val and Weidner 2014.

Notation: Some notation from graph theory is helpful. Let $V = V(n) = \{1, \dots, n\}$ denote a vertex set and define the corresponding directed edge set $E = E(n) = \{(v, v') : v, v' \in V(n), v \neq v'\}$. The vertices represent agents and the edges represent links. For a given link $e = (v, v')$, I refer to v as the sender and to v' as the receiver of the link. A graph g on V is a subset of E. For $g \subset E$, $(v, v') \in g$ is taken to mean that in g there is a directed link from v to v'. I use the terms network and graph interchangeably. For arbitrary graphs g, define the vertex function V that maps each graph g into the set of its constituent vertices. For a given graph g, the in-degree of agent i is defined as the number of links received by i, or $d_i^{\text{in}}(g) = \sum_{j \neq i} \mathbf{1}\big((j,i) \in g\big)$. Similarly, the out-degree of agent i is defined as the number of links sent by i, or $d_i^{\text{out}}(g) = \sum_{j \neq i} \mathbf{1}\big((i,j) \in g\big)$. The degree of agent i is the the sum of her in-degree and her out-degree.

3.2 The linking model

3.2.1 Model definition

Agent $i = 1, \ldots, n$ may link to any agent $j \neq i$. Linking decisions follow a static binary choice model. Consider the link $e = (i, j)$ and let Y_e denote a binary variable that is one if e is realized and zero otherwise. Sender i links to receiver j and $Y_e = 1$ if link surplus exceeds a link-specific shock,

$$Y_e = \mathbf{1}(Y_e^{\text{SP}} \geq \epsilon_e).$$

Y_e^{SP} is the latent link surplus and $(\epsilon_e)_{e \in E}$ is a vector of stochastically independent shocks with known distribution F. The assumption of independent surplus shocks precludes network externalities. For F any sufficiently smooth distribution can be chosen. Other authors require the shock distribution to be logistic (Holland and Leinhardt 1981; Graham 2014). For the link $e = (i, j)$ the latent surplus is given by

$$Y_{(i,j)}^{\text{SP}} = X'_{(i,j)}\theta^0 + \gamma_i^{S,0} + \gamma_j^{R,0}. \tag{3.1}$$

Here, $X'_{(i,j)}\theta^0$ is a measure of social distance between i and j based on observed characteristics and hence represents the homophily part of the utility function. The parameter θ^0 specifies homophily preferences and is unknown. The link-specific vector of observed covariates $X_{(i,j)}$ is typically a transformation of $(X_i, X_j, Z_{(i,j)})$, where X_k are observed characteristics of agent k and Z_e are edge-specific covariates. The covariate profile of the network is denoted by $\mathbf{X} = \{X_e : x \in E\}$.

The variables $\gamma_i^{S,0}$ and $\gamma_j^{R,0}$ are unobserved agent effects. Similar to Holland and Leinhardt 1981, the sender or productivity effect $\gamma_i^{S,0}$ encapsulates all aspects related to agent i's eagerness to initiate links to other agents. Similarly, the receiver or popularity effect $\gamma_j^{R,0}$ summarizes all of agent j's qualities that determine her attractiveness as a linking partner. In Section 3.6, I give an interpretation of the unobserved effects for a concrete example.

Sender and receiver effects are treated as fixed effects, allowing for arbitrary correlations between productivity, popularity and observed characteristics. Due to the fixed effects approach, agent effects may subsume unobserved determinants of linking behavior such as heterogeneous preferences or agent strategies in a latent game of social interaction. Since inference is conditional on unobserved agent effects, strategies can be arbitrarily correlated.

As in Holland and Leinhardt 1981, identification of the location of the unobserved effects is achieved by the normalization

$$\sum_{i \in V(n)} (\gamma_i^{S,0} - \gamma_i^{R,0}) = 0. \tag{3.2}$$

The specification of link surplus in (3.1) introduces three implicit assumptions. First, the three components homophily, productivity and popularity are required to be additively separable. This rules out, for example, linking behavior based on homophily preferences that change according to how popular a potential linking partner is. Note, however, that the separability assumption does not restrict correlations between the three components of link surplus. Secondly, it is assumed that the homophily component belongs to a known parametric family. Thirdly, all characteristics contributing to the homophily component are assumed to be observable to the econometrician.

The observability assumption is relaxed in latent space models (Hoff, Raftery, and Handcock 2002; Krivitsky et al. 2009). In these models, the mutual attraction between agents is allowed to depend on distance in a low-dimensional latent space. The class of latent space models does not, however, nest my model. The models in this class impose a relatively simple structure of unobserved heterogeneity that can make it impossible to correctly disentangle homophily from unobserved heterogeneity (Graham 2014).

To establish a baseline, I compare my linking model to a related model without fixed effects. For this model equation (3.1) is replaced by

$$Y^{\mathrm{SP}}_{(i,j)} = X'_{(i,j)}\theta^{H,0} + X'_i\theta^{S,0} + X'_j\theta^{R,0}, \tag{3.3}$$

where $\theta^{H,0}$, $\theta^{S,0}$ and $\theta^{R,0}$ parameterize productivity, attractiveness and homophily, respectively. For $e = (i,j)$, let $\mathcal{X}_e = (X'_e, X'_i, X'_j)'$ denote the variables predicting the generation of link e and let $\theta^{P,0} = (\theta^{H,0\prime}, \theta^{S,0\prime}, \theta^{R,0\prime})'$. As this model does not account for heterogeneity in a nonparametric way, it will be referred to as the *parametric model* in the remainder of the paper. The nonparametric specification for the sender effect γ^S_i is replaced by $X'_i\theta^{S,0}$. Similarly, the receiver effect γ^R_j is specified as $X'_j\theta^{R,0}$.

It is convenient to let $\pi_{(i,j)} = \gamma^{S,0}_i + \gamma^{R,0}_j$ denote the unobserved component of the surplus of link $e = (i,j)$. This way, equation (3.1) can be written more succinctly as $Y^{\mathrm{SP}}_e = X'_e\theta^0 + \pi_e$. Also, let $\gamma^S = (\gamma^S_1, \ldots, \gamma^S_n)'$, $\gamma^R = (\gamma^R_1, \ldots, \gamma^R_n)'$ and $\phi^0 = (\gamma^{S\prime}, \gamma^{R\prime})'$, and let $p_e = F(X'_e\theta^0 + \pi_e)$ denote the conditional probability of $Y_e = 1$. Throughout, $\bar{\mathbb{E}}$ denotes the expectation operator conditional on unobserved effects and the covariate profile, and E denotes the unconditional expectation operator.

3.2.2 Local structure

This section explores ramifications of the linking model for larger structures in the network by considering network relationships within triads (groups of three). I will focus on a triadic configuration called transitivity. Agents i, j and k are in a transitive relationship if, possibly upon reshuffling the labels within the triad, the network contains the links (i,j), (j,k) and (i,k). A tendency for transitive closure will result in a large number of links between connected nodes. In this regard, transitivity is a driver of local clustering.

To define measures of transitivity, let (i,j,k) denote a triple of distinct vertices. For a given graph g the triple is transitive if $\{(i,j),(j,k),(i,k)\} \subset g$. Figure 3.1 gives a visual

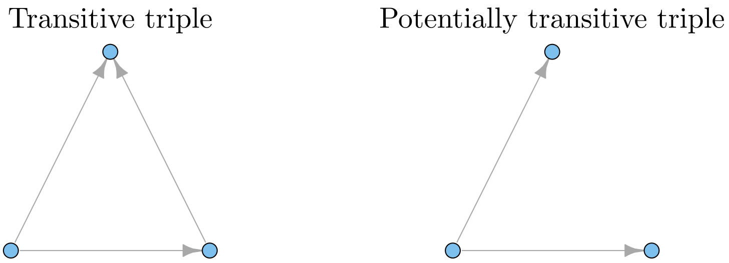


Figure 3.1: A transitive and a potentially transitive triple.

representation of a transitive triple. Define the set of all possible transitive triples[1]

$$\begin{aligned} B =& \{\beta \subset E(n) : \beta \text{ is a transitive triple}\} \\ =& \{\{(v_1, v_2), (v_2, v_3), (v_1, v_3)\} : \{v_1, v_2, v_3\} \subset V, |\{v_1, v_2, v_3\}| = 3\}. \end{aligned}$$

For every $\beta \in B$ take $\beta = \{\beta_1, \beta_2, \beta_3\}$, noting that the labelling of the edges is arbitrary. Let $T_\beta = Y_{\beta_1} Y_{\beta_2} Y_{\beta_3}$ denote the binary indicator that is one if β is realized and zero otherwise. Measures of transitivity are based on the count of transitive triples

$$S_n = \sum_{\beta \in B} T_\beta = \sum_{\beta \in B} Y_{\beta_1} Y_{\beta_2} Y_{\beta_3}.$$

The simplest approach is to normalize S_n by the number of all possible transitive triples $|B| = n^3$. This is the statistic discussed in this paper. It translates a concept for undirected networks considered in Karlberg 1997 to directed networks. A popular alternative is to normalize by the number of potentially transitive triples (Karlberg 1999; Jackson 2008, p. 37, see also the right panel in Figure 3.1). This yields the clustering coefficient

$$Cl_n = \frac{S_n}{\sum_{i \in V} \sum_{j \in V \setminus \{i\}} \sum_{k \in V \setminus \{i,j\}} Y_{(i,j)} Y_{(i,k)}}. \tag{3.4}$$

In Section 3.7, I indicate how my analysis can be extended to the clustering coefficient. Most of the time, I will drop the normalization and refer to S_n as realized or *observed transitivity*, and to its population counterpart $\bar{\mathbb{E}}\, S_n$ as *predicted transitivity*. The normalization is expendable when comparing networks with the same number of agents.

It is well known that to correctly describe the transitivity of a graph, it is important to account for degree heterogeneity (Karlberg 1999). In the context of my model, this means that ignoring the unobserved effects can vastly distort predicted transitivity. This is best illustrated by way of a simple example.

Example 3.1 Suppose that the set of agents can be partitioned into a set of normal agents with cardinality n° and a set of popular agents (the "attractors") with cardinality $n^\star$. Each edge to a normal agent has probability p°, and each edge to an attractor

[1] The set B coincides with the set of all transitive triples in the complete graph on n vertices $g^n = E(n)$.

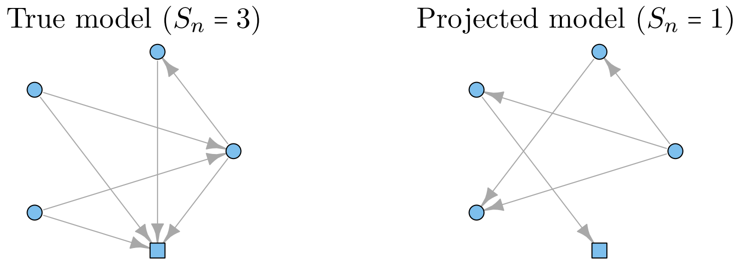


Figure 3.2: Realizations of the true model M^0 and the projected model M^{proj} from Example 3.1. The rectangle represents the attractor, circles represent normal agents.

has probability $p^\star$. Assume that popularity is the only relevant variable and that it can not be observed by the econometrician. Call this model M^0 and compare it to its projection M^{proj} onto the space of models that ignore popularity. In the projected model the common link probability is given by

$$p = \frac{n^\circ}{n} p^\circ + \frac{n^\star}{n} p^\star.$$

Now, adopt an asymptotic framework by considering a sequence of models M_n^0 and compare transitive triple counts between the true and the projected model. In the appendix it is shown that if $\frac{p^\circ}{p^\star} \to \alpha$, $0 \leq \alpha < 1$, and $\frac{n^\circ}{n^\star} \to \lambda > 0$ then

$$\frac{E_{M_n^0}[\# \text{ transitive triples}]}{E_{M_n^{\text{proj}}}[\# \text{ transitive triples}]} \to 1 + e(\alpha, \lambda),$$

where

$$e(\alpha, \lambda) = \frac{(1-\alpha)^2 \lambda}{(1 + \alpha\lambda)^2} > 0.$$

Plots of the function e are provided in Figure 3.6 in the appendix. Details on the calculations are in Appendix 3.C.1. Realizations of the models M^0 and M^{proj} for the parametrization $n^\star = 1$, $n^\circ = 4$, $p^\star = .8$ and $p^\circ = .2$ are depicted in Figure 3.2.

The stylized model shows that estimates of network transitivity based on a dyadic linking model that ignores unobserved heterogeneity can vastly understate the true amount of transitivity present in the network.

The stochastic network induced by a correctly specified dyadic model replicates the behavior of the observed network. In particular, under asymptotics that take the number of agents to infinity, observed transitivity S_n is consistent for predicted transitivity $\bar{\mathbb{E}} S_n$. A natural approach for checking the validity of the dyadic model is to test the equality of these two quantities. Using transitivity to evaluate model performance is well-motivated. The dyadic model competes with alternative models that allow for network externalities.

For some applications, evidence for agent preferences for transitive closure has been gathered (Leung 2014; Mele 2013). Thus, observing transitivity that surpasses the level predicted by the dyadic model indicates that the dyadic model should be abandoned in favor of a model that admits network externalities. To make this interpretation plausible, it is crucial to specify a reference model that can account for all drivers of transitivity that are permitted in a dyadic model. As argued above, this includes possibly unobserved sources of degree heterogeneity. In Section 3.4.4, I develop a transitivity test based on a feasible version of the test statistic

$$\tilde{T}_n = n^{-3}(S_n - \bar{\mathbb{E}}\, S_n).$$

The prediction $\bar{\mathbb{E}}\, S_n$ is derived from the dyadic linking model from equation (3.1) and can therefore account for degree heterogeneity.

The idea of testing a network model by considering its predictions for network features that are not targeted by the model was first explored in Holland and Leinhardt 1978. Karlberg 1999 also offers transitivity tests based on this paradigm. In his models, degree heterogeneity does not have a structural interpretation. Its effect on transitivity is eliminated by conditioning on the observed degree sequence. Karlberg 1999 does not provide a large sample theory for the test and uses a simulation procedure to compute critical values. My test statistic is asymptotically normal and approximate critical values can be computed from this asymptotic distribution.

In the following some additional notation will be convenient. Consider a transitive triple β. For given observed and unobserved agent characteristics, let $\rho_\beta = \prod_{\beta_j \in \beta} p_{\beta_j}$ denote the probability of β being observed. Conditional on the realization of link e, this probability is denoted

$$\rho_{-e}(\beta) = \prod_{\substack{\beta_j \in \beta \\ \beta_j \neq e}} p_{\beta_j}.$$

Also, define $\boldsymbol{\beta}_e^n = \frac{1}{n} \sum_{\beta \in B : \beta \ni e} \rho_{-e}(\beta)$.

3.3 Parameter estimation and incidental parameter bias

3.3.1 Conditional ML estimation

To estimate the linking model from equation (3.1) the agents effects are treated as additional parameters to be estimated. The maximum likelihood estimator $(\hat{\theta}, \hat{\phi})$ of the vector of structural parameters (θ^0, ϕ^0) maximizes a conditional likelihood criterion under a constraint that imposes the normalization from equation (3.2). Formally, $(\hat{\theta}, \hat{\phi})$ solves

$$\begin{aligned} \max_{\theta, \phi} \quad & \frac{1}{n} \sum_{(i,j) \in E(n)} \ell_{(i,j)}(X_{(i,j)}, \gamma_i^S, \gamma_j^R) \\ \textit{subject to:} \quad & \sum_{i \in V(n)} (\gamma_i^S - \gamma_i^R) = 0 \end{aligned} \tag{3.5}$$

with

$$\ell_{(i,j)}(X_{(i,j)},\gamma_i^S,\gamma_j^R) = Y_{(i,j)} \log F\big(X_{(i,j)}\theta + \gamma_i^S + \gamma_j^R\big) \\ + (1 - Y_{(i,j)}) \log\Big(1 - F\big(X_{(i,j)}\theta + \gamma_i^S + \gamma_j^R\big)\Big).$$

For the theoretical analysis it is convenient to impose the normalization indirectly by penalizing the likelihood rather than by optimizing under a constraint (Fernández-Val and Weidner 2014). Let $v = (\iota_n', -\iota_n')'$, with ι_n denoting an n-vector of ones. The following penalized program is equivalent to (3.5). For fixed $b > 0$

$$(\hat{\theta},\hat{\phi}) \in \arg\max_{\theta,\phi} \frac{1}{n}\left\{\sum_{(i,j)\in E(n)} \ell_{(i,j)}(X_{(i,j)},\gamma_i^S,\gamma_j^R) - b\big(v'\phi\big)^2\right\}. \qquad (3.6)$$

3.3.2 Asymptotic framework and incidental parameter bias

The asymptotic framework considered in this paper sends the number of agents n to infinity. The number of parameters estimated by the program (3.5) is increasing in n. For every agent that is added to the network two additional parameters, namely the agent specific sender and receiver effects, have to be estimated. This renders the maximum likelihood estimator non-standard and leads to an incidental parameter problem (Neyman and Scott 1948). In the context of the network model this means that certain parameters are estimated with a bias that is of the same order as the stochastic part of the estimator. Let μ denote a generic parameter of the model. In the remainder of the paper I will explicitly consider $\mu = \theta$ and $\mu = n^{-3}\,\bar{\mathbb{E}}\, S_n$. Let $\hat{\mu}^{\mathrm{ML}}$ denote the plug-in estimator of μ using the maximum likelihood estimates from (3.5), and let $V_\mu = \lim_{n\to\infty} \operatorname{var}\hat{\mu}^{\mathrm{ML}}$ denote its asymptotic variance. Similar to non-linear panel models (Hahn and Newey 2004; Fernández-Val and Weidner 2014) the estimator $\hat{\mu}^{\mathrm{ML}}$ has a representation

$$\hat{\mu}^{\mathrm{ML}} = \mu + n^{-1}\mathrm{bias}_\mu + n^{-1}\mathcal{N}\left(0,V_\mu\right) + o_p\left(n^{-1}\right),$$

where bias_μ is an unobserved deterministic term. Due to the presence of this bias term, confidence intervals based on the normal approximations may not be centered on the true parameter and tests may not hold their nominal level. The estimator $\hat{\mu}^{\mathrm{ML}}$ is, however, consistent.

In this paper I propose a procedure for analytical bias correction. I derive an explicit expression for the leading term of the asymptotic bias in terms of observed and estimable quantities. The bias can then be consistently estimated by plugging in the maximum likelihood estimates. Subtracting the estimated bias from the maximum likelihood estimator yields an estimator that is asymptotically normal and centered at the true value.

Network data is fundamentally different from sampled panel data. In a panel, it is a reasonable approximation to treat individuals as isolated clones of one generic agent. In an asymptotic thought experiment we can keep adding more and more independent copies of the same individual to the pool. As the pool grows larger, we eventually learn the covariate generating distribution. The thought experiment does not translate well to networks. Agents interacting in a network are typically not strangers. Networks are built on top of older social structures that have shaped agent characteristics in the past.

To address this concern, I will interpret all estimations as conditional on observed covariates and unobserved effects. This comes at a cost, as it renders me unable to answer some questions that might be of economic interest. I can answer the question "What is the expected transitivity in a network consisting of a given set of agents with a certain configuration of covariates and unobserved effects?" However, I am unable to quantify fluctuations in observed transitivity that are due to random perturbations of agent characteristics. This is because the asympotic framework does not allow me to learn the generating process for agent characteristics.

3.3.3 Alternatives to analytical bias removal

In panel models, procedures following a similar approach of analytical bias removal have been shown to work well in a variety of models (Hahn and Newey 2004; Hahn and Kuersteiner 2011; Fernández-Val 2009; Fernández-Val and Weidner 2014). The main drawback of this method is that it relies on an explicit expression for the asymptotic bias. Even small changes to the model set-up can have repercussions for the asymptotic bias approximation, forcing the researcher to re-do tedious derivations. Also, implementing the bias formula can be a time consuming and error prone process. It is tempting to try out methods that are less model specific, in that they are able to detect and remove the bias without the researcher having to specify what it looks like. In the context of panel methods, bootstrap and jackknife-based methods fulfil this requirement.

Bootstrap-based bias correction (Kim and Sun 2013) tries to replicate the estimation problem by re-sampling from the error distribution. This approach has been thoroughly explored in the research leading up to this paper. Most networks are rather sparse and linking probabilities tend to be very small. Exploratory simulations have shown that in this setting a naive bootstrap procedure can be numerically unstable, and can occasionally suggest corrections that vastly overstate the true bias. Developing a bootstrap procedure that can cope with a sparse network structure is an interesting avenue for future research.

Jackknife corrections (Hahn and Newey 2004; Dhaene and Jochmans 2010; Fernández-Val and Weidner 2014) assume that the estimation problem is scalable in the following sense. Parameter estimation is associated with an asymptotic bias that can be well-approximated by a constant divided by the sample size. If the estimation procedure is applied to a subset of the original data, the estimator admits a similar representation with the *same* constant.

Under the scalability assumption, the constant can be recovered by noting that the difference between estimates from a small and a large sample is a known multiple of the constant. In panel models the invariance of the constant is justified by laws of large numbers that rest on assumptions limiting the between-individuals and time dependence of individual characteristics. Such assumptions are much harder to justify in a network setting, as I discussed above. Even with generous independence assumptions on individual characteristics, the link-specific covariates will still exhibit a substantial amount of correlation. To see this, note that n individual-specific covariates are mapped into $n(n-1)$ link-specific covariates.

From an implementation point of view, jackknifing is a very attractive option for panels. The observations can be partitioned into two sets that can be interpreted as observations from two distinct panel models. Estimating the shorter panel models is cheap, since the panel model has already been implemented. In contrast, it is not possible to estimate dyadic network models from partitioning sets of link-specific observations. While this does not invalidate jackknife inference in networks, it certainly makes it less appealing.

3.4 Analytical correction for incidental parameter bias

3.4.1 Assumptions and notation

For convenience of notation we introduce some abbreviations. Let

$$F_{(i,j)} = F_{(i,j)}(X_{(i,j)}\theta^0 + \gamma_i^{S,0} + \gamma_j^{R,0})$$

denote the distribution function of the (i,j) observation evaluated at the true index and let $f_e = \partial F_e$ and $\partial f_e = \partial^2 F_e$ denote its first and second derivatives also evaluated at the true index. Let $H_e = f_e/(F_e(1-F_e))$ and $\omega_e = f_e H_e$.

Assumption 3.1 (Regularity conditions)

(i) The link function F is three times continuously differentiable.

(ii) Let $f_e^{(k)} = \partial^k f_e$, $k > 0$, and $f_e^{(0)} = f_e$. For all non-negative integers k_1, k_2 such that $k_1 + k_2 \leq 2$

$$\limsup_{n\to\infty} \max_{e\in E(n)} \left| \frac{f_e^{(k_1)} f_e^{(k_2)}}{F_e(1-F_e)} \right| < \infty.$$

Moreover,

$$\limsup_{n\to\infty} \max_{e\in E(n)} \frac{f_e}{F_e(1-F_e)} < \infty.$$

(iii) For a positive constant b_L and almost all $e \in E(n)$

$$\omega_e = \frac{f_e^2}{F_e(1-F_e)} \geq b_L.$$

(iv) The population version of the penalized objective function (3.6) is strictly concave.

Assumption 3.1 imposes the smoothness conditions from Assumption 4.1 in Fernández-Val and Weidner 2014 on the network model. Part (i) of the assumption requires the link function F to be sufficiently smooth. Popular choices such as the probit or the logit link satisfy the requirement. Item (ii) ensures that higher-order derivatives of the likelihood are well-behaved. In general, this assumption restricts both the shape of the link function and the distribution of the true latent indices. For some link functions, such as the one-dimensional Gaussian family, this assumption is satisfied for arbitrary index distributions. Item (iii) guarantees that the inverse of the penalized Hessian is well-behaved. This assumption is included primarily for technical convenience. For the probit model, it is satisfied if the latent index is bounded away from infinity. This in turn means that link probabilities are not allowed to vanish. In particular, it is not permitted to enforce asymptotic sparsity of the generated graph by letting the unobserved effects approach negative infinity. This might seem too restrictive. However, as I illustrate for a model without unobserved effects in Appendix 3.C.2, explicitly modeling sparsity does not require strong additional assumptions, nor does it change the analysis in a substantial way. Therefore, in practical applications and for a given sample size, the link function can be interpreted as incorporating the appropriate sparsity constant. Lastly, the concavity assumption (iv) ensures that — at least asymptotically — there is a unique solution to the program (3.5). It will typically be met if the parametric part of the model describes a symmetric distance measure and if there is sufficient between-individual variation in the observed covariates.

Fernández-Val and Weidner 2014 show that certain projections are helpful in describing the asymptotic bias. To define corresponding projections for the network model, let $P_\phi A$, for any $A = (A_e)_{e\in E}$, denote the orthogonal projection onto the space spanned by the fixed effects under an inner product weighted by $\omega_e^{1/2}$. In particular, $(PA)_{i,j} = \bar{\gamma}_i^S + \bar{\gamma}_j^R$ for any $(\bar{\gamma}_i^S, \bar{\gamma}_i^R)_{i\in V}$ solving

$$\min_{\gamma_i^S, \gamma_j^R} \sum_{i\neq j} \omega_{(i,j)} \left(A_{(i,j)} - \gamma_i^S - \gamma_j^R\right)^2 .$$

Let $\tilde{X}_e$ denote the component-wise residual of a projection of the X_e onto the space spanned by the fixed effects, i.e., for $k = 1, \dots, \dim(X_e)$ and $A = (X_{e,k})_{e\in E}$ let $\tilde{X}_{e,k} = X_{e,k} - \left(P_\phi A\right)_e$.

3.4.2 Inference on homophily parameter

We will first consider estimation of the homophily parameter θ. The following theorem gives the asymptotic distribution of $\hat{\theta}$, the maximum likelihood estimator of θ solving the program (3.5).

Theorem 3.1 (Estimation of homophily parameter) *Let*

$$\begin{aligned} B_\infty &= -\lim_{n\to\infty} \frac{1}{2n} \sum_{i\in V} \frac{\sum_{j:j\neq i} H_{(i,j)} \partial f_{(i,j)} \tilde{X}_{(i,j)}}{\sum_{j:j\neq i} \omega_{(i,j)}} \\ D_\infty &= -\lim_{n\to\infty} \frac{1}{2n} \sum_{j\in V} \frac{\sum_{i:i\neq j} H_{(i,j)} \partial f_{(i,j)} \tilde{X}_{(i,j)}}{\sum_{i:i\neq j} \omega_{(i,j)}} \\ \tilde{W}_\infty &= \lim_{n\to\infty} \frac{1}{n^2} \sum_{e\in E(n)} \omega_e \tilde{X}_e \tilde{X}_e'. \end{aligned}$$

Suppose that Assumption 3.1 holds and the above limits exist conditionally on $(\mathbf{X}, \phi^0)$ *and that* $\tilde{W}_\infty > 0$. *Then conditional on* $(\mathbf{X}, \phi^0)$

$$n(\hat{\theta} - \theta^0) \xrightarrow{d} \tilde{W}_\infty^{-1} B_\infty + \tilde{W}_\infty^{-1} D_\infty + \mathcal{N}(0, \tilde{W}_\infty^{-1}).$$

The theorem states that upon appropriate normalization the difference between the estimator and the true parameter is asymptotically normal and centered at $\text{bias}_\theta = \tilde{W}_\infty^{-1} B_\infty + \tilde{W}_\infty^{-1} D_\infty$. The asymptotic bias term is due to the unobserved effects that enter the estimation problem as an incidental parameter. The first term in the expression for the asymptotic bias can be attributed to the estimation of the sender effects and the second term can be attributed to the estimation of the receiver effects.

The rate of convergence to the limiting distribution is $O(n)$. Note that we observe $n(n-1)$ potential links. so that n behaves like the square root of the total number of link observations. Therefore, convergence is at the usual parametric rate (cf. Graham 2014).

Note that the theorem implies a version of the stochastic expansion sketched in Section 3.3.2. For bias_θ as defined above

$$\hat{\theta} = \theta^0 + n^{-1} \text{bias}_\theta + n^{-1} \mathcal{N}\big(0, \tilde{W}_\infty^{-1}\big) + o_p\left(n^{-1}\right).$$

To center the estimator at the true value we want to remove the second term in this expansion. Direct application of the theorem is infeasible as the asymptotic bias bias_θ is a function of the true latent index, which is unobserved. Since we have consistent estimators of θ^0 and ϕ^0 at our disposal, we can construct a consistent plug-in estimator of the asymptotic bias.

Define $\hat{\tilde{W}}_n^{-1}$, $\hat{B}_n$ and $\hat{D}_n$ as $\tilde{W}_\infty^{-1}$, B_∞ and D_∞, respectively, with the true latent index replaced by $X_e \hat{\theta} + \hat{\pi}_e$ and limits replaced by finite sums over the observed vertex set. Here, $\hat{\pi}_{(i,j)} = \hat{\gamma}_i^S + \hat{\gamma}_j^R$. The estimator with analytical bias correction is given by

$$\hat{\theta}^A = \hat{\theta} - n^{-1} \hat{\tilde{W}}_n^{-1} \hat{B}_n - n^{-1} \hat{\tilde{W}}_n^{-1} \hat{D}_n.$$

Theorem 3.1 is closely related to a result for the binary choice panel model from Example 1 in Fernández-Val and Weidner 2014. To see this more clearly, we need to explore the relationship between my network model and panel models.

First off, we have to think of each individual as occupying two distinct roles. For some links the individual will take on the role of the sender, and for other links it will take on the role of the receiver. Similar to certain arguments in game theory, this changes the setting from one where n agents interact to one where $2n$ agents interact. In the network model, two unobserved effects feed into the equation determining the linking behavior for link (i, j), namely, the sender effect of sender i and the receiver effect of receiver j. This is similar to a binary choice panel model with individual and time fixed effects. In the panel model, the binary choice of individual i in period t depends on two unobserved effects, namely, the individual effect of individual i and the time effect for period t. In this sense, an (i, j) observation in the network model maps to an (i, t) observation in the panel model. This relationship is obfuscated by the fact that senders and receivers in the network model share the same labels, whereas the individual and time dimensions in a panel model are labeled differently. The network model is, however, not completely congruent to the panel model. Note that self-links are not allowed. Therefore, sender i will meet all receivers $j \neq i$ but will never meet receiver i. This is different from the panel model where all individuals are observed at all time periods.

3.4.3 Inference on local structure

This section discusses estimation of predicted transitivity $\bar{\mathbb{E}}\, S_n$. For link formation, I consider the linking model with unobserved effects from equation (3.1), as well as the parametric model from equation (3.3).

Measuring features of local structure such as transitivity is a network-specific estimation problem with no counterpart in panel models. From a technical perspective, however, it is noted that predicted transitivity averages over structural parameters in a way that is reminiscent of a marginal effect in a panel model. This relationship can be exploited in the theoretical analysis.

To emphasize that the success probabilities for transitive triples are functions of the structural parameters and the observed covariate profile, I will write $\rho_\beta = \rho_\beta(\mathbf{X}, \phi^0, \theta^0)$ for transitive triples β when discussing the model with unobserved effects. The number of transitive triples predicted by the dyadic linking model is

$$\bar{\mathbb{E}}\, S_n = \sum_{\beta \in B(n)} \rho_\beta = \sum_{\beta \in B(n)} \rho_\beta(\mathbf{X}, \phi^0, \theta^0) = \sum_{\beta \in B} \prod_{e \in \beta} F_e(X_e \theta^0 + \pi_e^0).$$

A plug-in estimator of this quantity can be constructed by replacing the structural parameters by their maximum likelihood estimators,

$$\widehat{\mathrm{E}\, S_n} = \sum_{\beta \in B(n)} \rho_\beta(\mathbf{X}, \hat{\phi}, \hat{\theta}^0). \tag{3.7}$$

Let $D_e^n = f_e \boldsymbol{\beta}_e^n$ and

$$R_{\theta,n} = \frac{1}{n^3} \sum_{\beta \in B} \partial_\theta \rho_\beta = \frac{1}{n^2} \sum_{e \in E(n)} D_e^n X_e.$$

For the model without unobserved effects I adopt similar notation.

As in the discussion of the homophily parameter, all inference will be conditional on unobserved effects and the observed covariate profile. For the limiting distribution to be well-defined, certain limits will be required to exist. For the parametric model, I will investigate the plausibility of this assumption by providing conditions on the data generating process that guarantee that the required limits exist. I conjecture that similar arguments can be made for the model with unobserved effects. Consider the following assumption about the data generating process for the covariates.

Assumption 3.2
The $\mathcal{X}_e$, $e \in E(n)$, are identically distributed and

$$V(e) \cap V(e') = \varnothing \;\Rightarrow\; \mathcal{X}_e \perp\!\!\!\perp \mathcal{X}_{e'}.$$

Moreover, the components of $\mathcal{X}_e$ have bounded fourth moments.

To interpret this assumption, recall that the function V returns the vertices of a graph so that for $e = (i, j)$ we have $V(e) = \{i, j\}$. The assumption restricts the dependence between edge-specific covariates. As discussed above, it is not appropriate to assume full independence of the edge-specific covariates. Assumption 3.2 offers a substantially weaker alternative by requiring independence of covariates only for edges that have no common vertices.

The following result characterizes the asymptotic distribution of the estimator of predicted transitivity in the model without unobserved effects.

Theorem 3.2 (Predicted transitivity without unobserved effects) *Consider the model without unobserved effects from equation* (3.1). *Suppose that the link function F is bounded away from zero and one on the support of the latent index, and that it is three times continuously differentiable. Let $R_{\theta,\infty} = \lim_{n\to\infty} R_{\theta,n}$,*

$$W_\infty = \lim_{n\to\infty} \frac{1}{n^2} \sum_{e\in E(n)} \omega_e \mathcal{X}_e \mathcal{X}_e',$$

$$V_T^{(a)} = \lim_{n\to\infty} \frac{1}{n^2} \sum_{e\in E(n)} \omega_e \left\{ \left(R_{\theta,\infty}\right)' W_\infty^{-1} \mathcal{X}_e \right\}^2 .$$

Suppose that conditional on $\mathbf{X}$ *all limits exist and that* $V_\theta^{(a)} > 0$. *Then*

$$n^{-2} \left(\widehat{\mathrm{E}\, S_n} - \bar{\mathbb{E}}\, S_n \right) \xrightarrow{d} \mathcal{N}\left(0, V_T^{(a)}\right)$$

conditionally on $\mathbf{X}$. *If Assumption 3.2 holds, then on a set with probability approaching one* $R_{\theta,\infty}$, W_∞ *and* $V_T^{(a)}$ *exist. In particular,* $R_{\theta,\infty} = \mathrm{E}\left[\partial_\theta \rho_\beta(\mathbf{X}, \theta^0)\right]$ *and* $W_\infty = \mathrm{E}\left[\omega_e \mathcal{X}_e \mathcal{X}_e'\right]$.

Remark 3.1 The assumption that p_e is bounded away from zero is undesirable in a network context. It will lead to networks that are asymptotically dense. An analogue result for a model with link function $F_n = a_n^{-1} F$ depending on n can be found in the Appendix. Here, a_n is a known deterministic sequence. The main restriction is that $a_n^{-1} n^2 \to \infty$. This assumption is not too strong. In particular, it allows for degree sequences that are bounded away from infinity.

It should be noted that sometimes $\bar{\mathbb{E}}\, S_n$ might not be the right quantity to estimate. For applications such as comparing transitivity across different networks, the unconditional mean $\mathrm{E}\, S_n$ is more informative. Under appropriate conditions on the sampling process of the covariates and the unobserved effects, $\widehat{\mathrm{E}\, S_n}$ consistently estimates $\mathrm{E}\, S_n$. However, $V_T^{(a)}$ given in Theorem 3.2 will not capture the true variance of $\widehat{\mathrm{E}\, S_n}$ as an estimator of $\mathrm{E}\, S_n$ as it fails to take into account fluctuations of $\bar{\mathbb{E}}\, S_n$ around $\mathrm{E}\, S_n$ as a source of variation. Under common specifications of the data generating process, these fluctuations can dominate the asymptotic distribution, rendering parameter estimation asymptotically negligible (cf. Fernández-Val and Weidner 2014).

In this paper, I focus on transitivity for a given set of agents, and on testing predicted transitivity against observed transitivity. For these purposes, $\bar{\mathbb{E}}\, S_n$ is an appropriate measure.

To present the companion result to Theorem 3.2 for the model with unobserved effects, it is convenient to introduce new notation. We will need certain derivatives of predicted transitivity with respect to sender and receiver effects. Let

$$\delta_i^S = n\left(\partial_{(\gamma_i^S)^2}\frac{1}{n^3}\sum_{\beta\in B}\rho_\beta\right) \quad \text{and} \quad \delta_j^R = n\left(\partial_{(\gamma_j^R)^2}\frac{1}{n^3}\sum_{\beta\in B}\rho_\beta\right).$$

Also, note that

$$\partial_{\gamma_i^S}\left(\frac{1}{n^3}\sum_{\beta\in B}\rho_\beta\right) = \frac{1}{n^2}\sum_{j:j\neq i} D^n_{(i,j)}$$

and that a corresponding equation holds for derivatives with respect to receiver effects. For $A = (-D_e^n/\omega_e)_{e\in E}$ and P_ϕ defined as above let $\Psi_e = \left(P_\phi A\right)_e$.

Theorem 3.3 (Predicted transitivity with unobserved effects) *Consider the model with unobserved effects from equation* (3.1). *Let*

$$\Xi_n = \frac{1}{n^2}\sum_{e\in E} D_e^n \tilde{X}_e$$

and $\Xi_\infty = \lim_{n\to\infty}\Xi_n$. *For* B_∞, D_∞ *and* $\tilde{W}_\infty$ *as defined in Theorem 3.1 let*

$$B_\infty^{TT} = \Xi_\infty' \tilde{W}_\infty^{-1} B_\infty + \lim_{n\to\infty}\frac{1}{2n}\sum_i \frac{\sum_{j:j\neq i}\left(\delta_i^S + H_{(i,j)}\Psi_{(i,j)}\partial f_{(i,j)}\right)}{\sum_{j:j\neq i}\omega_{(i,j)}} \qquad \textit{and}$$

$$D_\infty^{TT} = \Xi_\infty' \tilde{W}_\infty^{-1} D_\infty + \lim_{n\to\infty}\frac{1}{2n}\sum_j \frac{\sum_{i:i\neq j}\left(\delta_j^R + H_{(i,j)}\Psi_{(i,j)}\partial f_{(i,j)}\right)}{\sum_{i:i\neq j}\omega_{(i,j)}}.$$

Let

$$V_T = \lim_{n\to\infty}\frac{1}{n^2}\sum_e \omega_e\left\{\Xi_\infty'\tilde{W}_\infty^{-1}\tilde{X}_e - \Psi_e\right\}^2.$$

Assume that, conditional on $(\mathbf{X},\phi^0)$, *Assumption 3.1 holds, all limits are well defined and finite, and* $V_T > 0$. *Conditional on* $(\mathbf{X},\phi^0)$

$$n^{-2}\left(\widehat{\mathrm{E}\,S_n} - \bar{\mathbb{E}}\,S_n\right) \xrightarrow{d} B_\infty^{TT} + D_\infty^{TT} + \mathcal{N}(0, V_T).$$

Remark 3.2 If $\hat{\theta}$ in equation (3.7) is replaced by the bias corrected estimator $\hat{\theta}^A$ the respective first term in the expression for B_∞^{TT} and D_∞^{TT} drops out. This is similar to a corresponding result for marginal effects in Fernández-Val and Weidner 2014.

Theorem 3.3 shows that the plug-in estimator of predicted transitivity is affected by incidental parameter bias. The first component of the asymptotic bias, B_∞^{TT}, is due to the estimation of the sender effects, and the second component of the bias, D_∞^{TT}, is due to the estimation of the receiver effects.

Note that $\bar{\mathbb{E}}\,S_n$ is of the same order as n^3, so that convergence is, again, at the parametric rate n.

As before, the expression for the asymptotic bias offers a recipe for analytical bias correction. Define $\hat{B}_n^{TT}$ and $\hat{D}_n^{TT}$ as B_∞^{TT} and D_∞^{TT}, respectively, with the true latent index replaced by $X_e\hat{\theta} + \hat{\pi}_e$ and limits replaced by finite sums over the observed vertex set. The bias corrected estimator is given by

$$\widehat{\mathrm{E}\,S_n}^A = \widehat{\mathrm{E}\,S_n} - n^2\hat{B}_n^{TT} - n^2\hat{D}_n^{TT}.$$

3.4.4 Testing local structure

This section formalizes the test idea developed in Section 3.2.2. In $\tilde{T}_n$, replace $\bar{\mathbb{E}}\,S_n$ by its estimator $\widehat{\mathrm{E}\,S_n}$ to arrive at the feasible test statistic

$$T_n = n^{-3}\left(S_n - \widehat{\mathrm{E}\,S_n}\right).$$

The transitivity test rejects for large values of the test statistic. As null models I will consider both the model with and the model without unobserved effects.

Theorem 3.4 (Testing transitivity without unobserved effects) *Consider the model without unobserved effects from equation* (3.1) *and suppose that the conditions of Theorem 3.2 are satisfied. Moreover, let*

$$V_S^{(a)} = \lim_{n\to\infty}\frac{1}{n^2}\sum_{e\in E(n)} F_e(1-F_e)\left\{\boldsymbol{\beta}_e^n - H_e\left(R_{\theta,\infty}\right)' W_\infty^{-1}\mathcal{X}_e\right\}^2.$$

Suppose that, conditional on $\mathbf{X}$*,* $V_S^{(a)}$ *exists and that* $V_S^{(a)} > 0$*. Then*

$$nT_n = n^{-2}\left(S_n - \widehat{\mathrm{E}\,S_n}\right) \xrightarrow{d} \mathcal{N}\left(0, V_S^{(a)}\right).$$

conditional on $\mathbf{X}$.

Remark 3.3 In the Appendix it is shown that $V_S^{(a)}$ can be replaced by

$$V_S^{(b)} = \lim_{n\to\infty}\frac{1}{n^2}\sum_{e\in E(n)}\sum_{\substack{\beta,\beta' \\ \beta\cap\beta'=\{e\} \\ |V(\beta)\cap V(\beta)|=2}}\frac{(\rho_{-e}(\beta) - \frac{1}{3}X_e^\Diamond)(\rho_{-e}(\beta') - \frac{1}{3}X_e^\Diamond)}{n^2}F_e(1-F_e)$$

where $X_e^\Diamond = H_e\left(R_{\theta,\infty}\right)' W_\infty^{-1}\mathcal{X}_e$. This shows that the asymptotic variance is a function of all subgraphs that are formed by taking two transitive triples that share exactly two vertices and one edge. Note that this representation of the variance is not well suited for computational purposes as compared to $V_S^{(a)}$ it increases computational complexity from $O(n^3)$ to $O(n^4)$.

For a brief heuristic description of how to derive the asymptotic distribution of the test statistic, write

$$nT_n = -n^{-2}(\widehat{\mathrm{E}\,S_n} - \bar{\mathbb{E}}\,S_n) + n^{-2}(S_n - \bar{\mathbb{E}}\,S_n).$$

For the first term we can exploit a stochastic expansion derived in the proof of Theorem 3.2. Characterizing the second term is related to deriving the asymptotic distribution of the triad census in the analysis of the original τ^2-test (Holland and Leinhardt 1978). In seminal work, Holland and Leinhardt 1970 and Holland and Leinhardt 1976 give an explicit formula for the variance under their choice of reference distribution and conjecture asymptotic normality. To date, I am unaware of a formal statement supporting this conjecture. My proof of asymptotic normality exploits similarities between the count of transitive triples and a certain class of U-statistics. For many reference distributions, the distributional analysis of the triad census is amendable to the same approach.

In the model with unobserved effects we have to account for incidental parameter bias.

Theorem 3.5 (Testing transitivity with unobserved effects) *Consider the model with unobserved effects from equation* (3.1)*. Suppose that the conditions of Theorem 3.3 are satisfied. Moreover, let*

$$\upsilon_e = \boldsymbol{\beta}_e^n - H_e\left(\Xi_\infty' \tilde{W}_\infty^{-1} \tilde{X}_e - \Psi_e\right)$$

and

$$V_S = \lim_{n\to\infty} \frac{1}{n^2} \sum_e F_e(1-F_e)\upsilon_e^2.$$

Assume that, conditional on $(\mathbf{X}, \phi^0)$*,* V_S *exists and* $V_S > 0$*. Conditional on* $(\mathbf{X}, \phi^0)$

$$nT_n = n^{-2}\left(S_n - \widehat{\operatorname{E} S_n}\right) \xrightarrow{d} -B_\infty^{TT} - D_\infty^{TT} + \mathcal{N}(0, V_S).$$

This result can be used to construct a bias corrected test statistic

$$T_n^A = n^{-3}\left(S_n - \widehat{\operatorname{E} S_n}\right) + n^{-1}\hat{B}_n^{TT} + n^{-1}\hat{D}_n^{TT}.$$

The bias corrected test statistic is asymptotically centered at zero, and critical values can be computed from the normal distribution with variance V_S.

In Section 3.2.2, I pointed out a useful relationship between predicted transitivity and marginal effects in panel models. It is worth mentioning that the similarities do not extend to the testing problem. Marginal effects are properties of the population model that do not correspond to directly observable quantities. Therefore, they do not lend themselves to tests of model specification in the same way that predictions for local network structure do.

3.5 Simulations

In this section I report simulations that investigate the finite-sample performance of the analytical bias correction both for the estimator of the homophily parameter as well as for the estimator of predicted transitivity.

Agents $i = 1, \ldots, n$ are characterized by independent draws from the joint distribution of $(X_i, \gamma_i^S, \gamma_i^R)$. Here X_i is an agent-specific observed covariate distributed according to a $Beta(2,2)$ distribution (cf. the specification in Graham 2014). This distribution will

			bias		CI coverage	
n	λ	c	C	NC	C	NC
80	0.0	1.5	0.09	0.51	97.2	92.4
		1.7	0.03	0.52	96.8	92.6
	0.5	1.5	0.02	0.45	95.8	93.2
		1.7	−0.01	0.51	96.6	92.8
50	0.0	1.5	−0.01	0.41	96.2	94.4
		1.7	0.05	0.54	97.0	91.6
	0.5	1.5	−0.00	0.43	95.6	92.6
		1.7	0.21	0.75	97.0	90.0

Table 3.1: Simulation results for the homophily parameter θ^0. Columns labeled C refer to the bias-corrected estimator and columns labeled NC refer to the uncorrected estimator. The bias is in terms of the standard error of the estimator and the nominal level of the confidence interval is $1 - \alpha = 95\%$. Results are reported for $B = 500$ simulations.

endow a majority of agents with similar characteristics and concentrates deviations from the network average in a small, heterogeneous group of agents. This imitates a similar pattern observed in the application. The unobserved effects are generated according to

$$\gamma_i^S = \lambda(X_i - c) + (1-\lambda)(Beta(0.5, 0.5) - c) \quad \text{and}$$
$$\gamma_i^R = \lambda(X_i - c) + (1-\lambda)(Beta(0.5, 0.5) - c),$$

where the two *Beta* distributions are independent. The parametrization of the *Beta* distributions concentrates probability mass at the boundaries of the unit interval. This results in individuals being clustered into groups with low and high unobserved effects, similar to what is observed in the application. The parameter $\lambda \in (0,1)$ controls correlation between unobserved heterogeneity and observed attributes and the positive constant c shifts the success probability. In the simulations rather large values of c are chosen to emulate the small linking probabilities encountered in practice. For $e = (i,j)$ the link-specific homophily variables is given by $X_e = |X_i - X_j|$. Note that with this specification the $(X_e)_{e \in E}$ are not independent but Assumption 3.2 is satisfied for $\mathcal{X}_{(i,j)} = (X'_{(i,j)}, \gamma_S^S, \gamma_j^R)'$. The true value of the homophily parameter is $\theta^0 = 1.5$ and the link-specific disturbance is standard normally distributed.

Table 3.1 summarizes the behavior of the corrected and the uncorrected estimator of the homophily parameter in $B = 500$ simulations for different parameter values and two sample sizes. It reports the bias of the estimator in terms of its standard error as well as the empirical coverage of a confidence interval with nominal level $1 - \alpha = 95\%$. For the uncorrected estimator we observe a positive bias roughly the size of half a standard deviation. The bias is very effectively removed by the analytical bias correction, resulting in parameter estimates that are centered around the true value. This shows that, even in finite samples, bias correction based on an asymptotic approximation can be a powerful tool for increasing the precision of the estimates. Confidence intervals for

			bias			CI coverage			$\widehat{\mathbb{E}S_n}$		
n	λ	c	C	NC	P	C	NC	P	C	NC	P
80	0.0	1.5	0.08	1.49	−3.31	92.2	64.8	0.0	501	583	303
		1.7	0.05	1.91	−2.21	96.4	51.4	0.0	73	104	38
	0.5	1.5	0.06	1.66	−1.03	95.8	61.0	0.0	216	265	179
		1.7	0.14	2.20	−0.67	97.4	33.4	0.0	25	39	19
50	0.0	1.5	0.01	1.49	−2.00	88.6	63.4	0.0	125	152	77
		1.7	−0.19	1.69	−1.41	94.8	60.0	0.0	18	30	10
	0.5	1.5	0.07	1.72	−0.54	94.0	58.4	0.0	54	76	45
		1.7	−0.12	2.02	−0.46	95.0	43.0	0.4	6	12	5

Table 3.2: Simulation results for the estimator of predicted transitivity. Columns labeled C refer to the bias-corrected estimator, columns labeled NC refer to the uncorrected estimator and columns labeled P refer to the estimator based on the parametric model. Bias is reported in terms of standard deviations and the nominal level of the confidence interval is $1 - \alpha = 95\%$.

the uncorrected estimator are slightly undersized. After analytical bias correction the coverage probabilities are fairly close to the nominal size. The improvement is, however, not as substantial as it is for the bias.

Turning to predicted transitivity, I will also consider an estimator for the parametric model from equation (3.3). For the link $e = (i, j)$, the parametric model uses the observed covariates X_i and X_j to approximate sender and receiver effects. It is obvious from the specification of the data generating process that for $\lambda \neq 1$ this approximation will be imperfect.

Table 3.2 reports simulation results for three estimators of predicted transitivity. The estimates from the parametric model severely understate transitivity. This confirms the theoretical considerations from Section 3.2.2, showing that failure to account for unobserved sources of degree heterogeneity can result in severely down-biased transitivity estimates. Note that confidence intervals constructed from estimates based on the parametric model almost never contain the true parameter.

The fixed-effects estimator without bias correction exhibits a positive bias of about one-and-a-half to slightly over two standard deviations. Confidence intervals constructed from uncorrected estimates cover the true parameter with probability less than two-thirds. This is a substantial deviation from the nominal level of 95%. In the designs with low linking probability ($c = 1.7$), empirical coverage is as low as 30-40% in some cases.

In this simulation design, the analytical correction has very favorable finite sample properties. It picks up the bias almost completely. After applying the correction formula, the remaining bias is but a small fraction of a standard deviation. This considerably improves the normal approximation. The empirical coverage of confidence intervals computed from the asymptotic distribution is now very close to the nominal level of 95%.

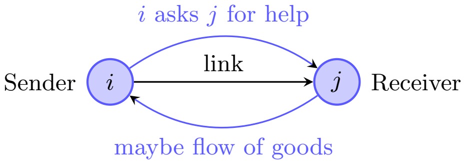


Figure 3.3: Definition of link: There is a link from i to j if, under a hypothetical situation, i would go to j to ask for help.

3.6 Application: Favor networks in Indian villages

I use the Indian village data from Banerjee et al. 2013 and Jackson, Rodriguez-Barraquer, and Tan 2012. This data set contains survey data from 75 Indian villages. In each village, about 30 - 40% of the adult population were handed out detailed questionnaires that elicited network relationships to other people in the same village as well as a wide range of socio-economic characteristics.

For this application, networks are defined on the village level. Therefore, the data set contains 75 network observations. For each village, the set of agents is given by the surveyed villagers. Links are defined by a social relationship related to anticipated favor exchanges.

In the presentation of my estimation results for the homophily component I will only consider a single village. To investigate the level of transitivity predicted by different dyadic models, I take advantage of the full data set and use all villages.

The directed network considered in this application is constructed from the survey questions "If you suddenly needed to borrow Rs. 50 for a day, whom would you ask?" and "If you needed to borrow kerosene or rice, to whom would you go to?". To set up the network, I let every surveyed individual send directed links to each of the individuals nominated in one of the two questions, provided that the nominee was also included in the survey. The network generated in this way is defined to be the network of interest. This avoids identification issues that arise when using a partial sample for inference on an imperfectly observed population network (Chandrasekhar and Lewis 2011). Addressing such problems is beyond the scope of this paper.

A link from agent i to agent j indicates that, in times of need, i would ask j for help. Note that, if j accedes to the request, the direction of the flow of goods will be opposite to the direction of the link. Figure 3.3 illustrates the behavior of two linked villagers under the hypothetical situation from the survey question.

It is instructive to discuss the significance of productivity, popularity and homophily in the context of this application. When deciding about whether to establish a link to some agent j, a sender i ponders whether j is able and willing to grant the request. Agent j's ability to provide help is affected by her own wealth and liquidity as well as i's ability to repay the loan or return the favor in the future. In the context of my model, the first effect contributes to j's popularity, and the second effect adds to i's productivity. Agent j's willingness to help is a function of how altruistic she is, of i's skill in negotiating the favor, and of how sympathetic j is towards i's plight. The first two considerations are, again, subsumed in j's popularity and i's productivity, respectively. It is plausible to assume that j is more sympathetic towards i the more similar the two of them are. This tendency is a manifestation of homophily. For example, j might have a high willingness to offer assistance to members of her own family, and have little inclination to help out individuals assigned to a different caste.

In the highly stylized decision model sketched in the previous paragraph, many drivers of productivity and popularity such as an innate predisposition towards acts of altruism, or expectations about future liquidity are inherently unobservable. In the dyadic linking model these unobserved factors will be captured by the unobserved agent effects. If the network is based on survey data, the sender effect can also subsume reporting behavior. This makes the estimator of the homophily parameter robust to some common forms of measurement error. The taste for homophily is captured by the parametric part of the latent index and it is assumed that all drivers of homophily are observed.

The fundamental assumption at the heart of the dyadic linking model is that for all linking decisions the dyad (or pair) is the relevant point of reference. This is an exogeneity assumption under which individuals evaluate each link in isolation of all other links. In particular, they do not care about the future network positions of their potential linking partners. In the context of the favor network this assumption is compelling for two reasons. First, as the network is based on a hypothetical, it is and remains largely unobserved, which makes it hard for individuals to condition their actions on network realizations. Secondly, the hypothetical transfer of goods that defines the network relation only affects the individuals within the dyad. This stands in stark contrast to other network relations, such as friendship networks, where it is natural to assume that individuals derive utility from links between their friends.

In other work the exogeneity assumption has been challenged. Jackson, Rodriguez-Barraquer, and Tan 2012 argue that reciprocation of favors is best enforced by the threat of other agents in the network to withhold future favors from shirking individuals. Leung 2014 provides estimates for preferences for local structure in favor networks. Since his model does not allow for unobserved sources of degree heterogeneity it is, however, hard to say whether the estimated effects are genuine or spurious (cf. Section 3.2.2). I will maintain the exogeneity condition as a working assumption. Below, I use the model specification test developed in this paper to critically assess its validity.

I now present detailed results for village 60, the largest village in the sample ($n = 414$). The estimation is based on the dyadic linking model with unobserved effects developed in this paper.

Table 3.3 lists all variables that are used in the specification for the homophily component. For the variables related to education, individuals are sorted into one of three bins according to their reported years of formal schooling. Individuals are assigned to the bin "SSLC" if they have obtained a Secondary Schooling Leaving Certificate. In India, this certificate is awarded to students who pass an examination at the end of grade 10. It is a prerequisite for enrolling in pre-university courses. All other individuals are assigned to "no education" if they have completed less than five years of schooling, and to "primary education" if they report at least five years of schooling. For caste membership I adopt the fairly broad categorization from the data set. Individuals are described as

Variable	Description
same caste	i and j belong to the same caste
age difference	absolute value of age difference between i and j
same family	i and j belong to the same family
same latrine	i and j both live in a house with an own latrine
same status	both i and j are household heads
same gender	i and j have the same gender
same village native	both i and j were born in the village
educ None-Primary	one of i and j has no education, the other has finished primary education
educ None-SSLC	one of i and j has no education, the other has a obtained a SSL certificate
educ Primary-SSLC	one of i and j has finished primary education, the other has obtained a SSL certificate

Table 3.3: Description of variables measuring homophily (X_e).

members of scheduled tribes, scheduled castes, other backwards castes (OBC's) or general castes.

	Coef	se	T	p-value
same caste	0.80	0.0484	16.44	0.0000
age difference	-0.01	0.0022	-5.97	0.0000
same family	2.45	0.0943	26.01	0.0000
same latrine	0.07	0.0331	1.97	0.0486
same status	0.05	0.0467	1.05	0.2921
same gender	0.53	0.0483	11.02	0.0000
same village native	0.04	0.0351	1.09	0.2735
educ None-Primary	-0.10	0.0428	-2.38	0.0173
educ None-SSLC	-0.19	0.0504	-3.82	0.0001
educ Primary-SSLC	-0.10	0.0499	-2.07	0.0388

Table 3.4: Homophily estimates for village 60.

Table 3.4 reports bias-corrected estimates and standard errors for the homophily component. Family ties are a dominating factor for determining targets for favor requests. This reflects a strong sense of solidarity between family members. Same caste membership and same gender are other strong determinants of the network relation. This is in line with findings in Leung 2014 who studies similar favor networks. The “same latrine” dummy, which is included as a proxy of similarities in wealth, has a comparably small estimated effect that is significant at the five percent but not at the one percent level. This indicates that the aversion to connecting to members of other castes is not driven solely by economic disparities. The education dummies are jointly significant at the one percent level (p-value = .0003). The estimated effect is almost linear, with a difference in education levels corresponding to one bin, decreasing the link surplus by roughly 0.1 points.

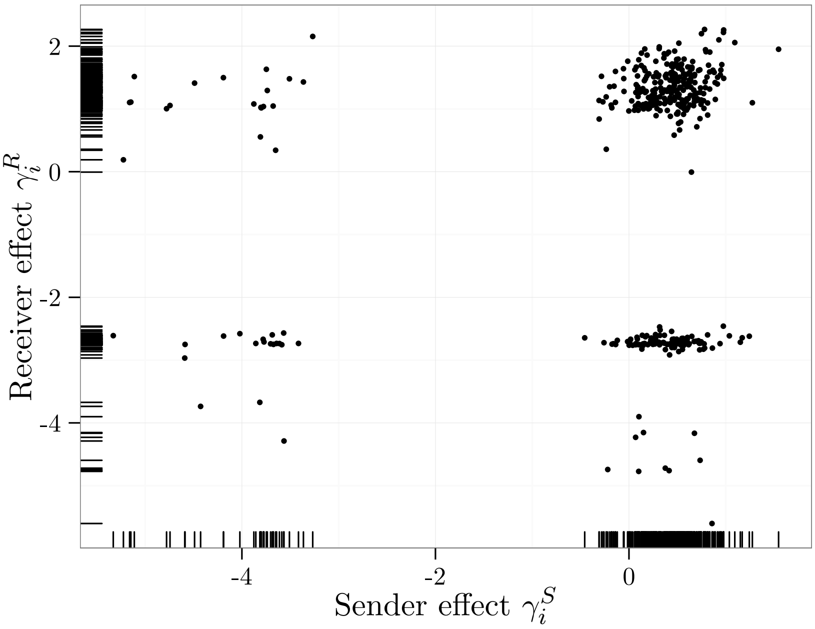


Figure 3.4: Unobserved heterogeneity in village 60. Unobserved types for agents $i = 1, \ldots, n$.

The unobserved type of agent i corresponds to the tupel (γ_i^S, γ_i^R). Thus, every agent type can be represented as a point on a two-dimensional plane. A plot of estimated types is provided in Figure 3.4 with sender and receiver effects centered at their common empirical mean[2]. The graph reveals an interesting pattern of unobserved heterogeneity. Types cluster into four distinct groups. The largest cluster consists of agents with relatively large sender and receiver effects (attractor-producers). The second largest cluster is composed of agents with relatively large sender effects and relatively small receiver effect (producers). The set of agents with below average sender effects splits neatly into a group with relatively large receiver effects (attractors) and a group with relatively small receiver effects (isolates).

This clustering pattern has interesting implications. First, there is no monotone relationship between sender and receiver effects. This suggests that productivity and popularity are distinct phenomena rather than two manifestations of one underlying variable such as social skill. This exemplifies the value of using data on the direction of links. Models for directed networks, such as Graham 2014, are by necessity restricted to modelling one-dimensional types and can therefore not reflect as rich a picture of the unobserved heterogeneity. Secondly, as most agents belong to clusters with large sender effects, unobserved heterogeneity will drive linking behavior mainly through variations in

[2] Note that the normalization from equation (3.2) imposes equality of the empirical mean of the sender effects and the empirical mean of the receiver effects.

receiver popularity. Sender productivity plays a less defining role.

The clusters can be compared along a wide range of observed characteristics such as age profiles (Figure 3.9). In the clusters with below-average receiver effects young and old people are over-represented, whereas individuals in their prime working age are under-represented. In the clusters with above-average receiver effects the pattern is inverted. About 12% of the agents in the attractor-producer cluster participate in self-help groups (SHGs). This is contrasted by almost non-existent participation rates in the other clusters. SHGs are savings and loan clubs organized at the village level. They might be related to productivity and popularity by attracting wealthier or more entrepreneurial villagers who are interested in depositing savings or taking out loans. Additional comparisons of cluster characteristics are provided in Table 3.6 in the appendix.

Unobserved agent effects determine in a fundamental way which links are formed. In Figure 3.7 and Figure 3.8 in the appendix, unobserved types are plotted against observed in-degrees and observed out-degrees, respectively. Agents belonging to the clusters with low receiver effects do not attract any links, and agents belonging to the clusters with low sender effects do not nominate any linking partners.

I now turn to estimating predicted transitivity and testing it against realized transitivity. To this end, I compare the model with unobserved effects to a benchmark given by the parametric model from equation (3.3). The parametric model approximates agent productivity and popularity using a rich array of observed characteristics detailed in Table 3.7 in the appendix. Results for almost all villages[3] in the dataset are summarized in Table 3.5 in the appendix.

For the model with unobserved effects, bias-corrected estimates are larger than the uncorrected estimates. On average, the size of the correction is about two-and-a half standard deviations. The magnitude of the estimated bias implies that for this application the bias correction is an essential part of the testing procedure. Failure to implement the correction will lead to substantially different test results.

As argued in Section 3.2.2, a model that does not account for all determinants of productivity and popularity will understate transitivity. The transitivity estimates from the parametric model are substantially lower than those obtained from the model with fixed effects, capturing on average only roughly 12% of the transitivity estimated by the model with unobserved heterogeneity. The degree to which the two estimates diverge suggests that unobserved heterogeneity plays a substantial role in driving degree heterogeneity. In other words, agent productivity and popularity are not explained well by observed characteristics.

The vast differences between the two models in terms of estimated level of transitivity are also reflected in the transitivity test. Figure 3.5 plots values of the test statistic for both models. The choice of model crucially affects the statistical significance of the difference between observed and predicted transitivity.

The parametric model rejects the null hypothesis of correct model specification for all villages (significance level $\alpha = 0.05$). In contrast, for the model with unobserved effects

[3] Some smaller villages for which collinearity issues in the specification of the parametric model arise, have been excluded.

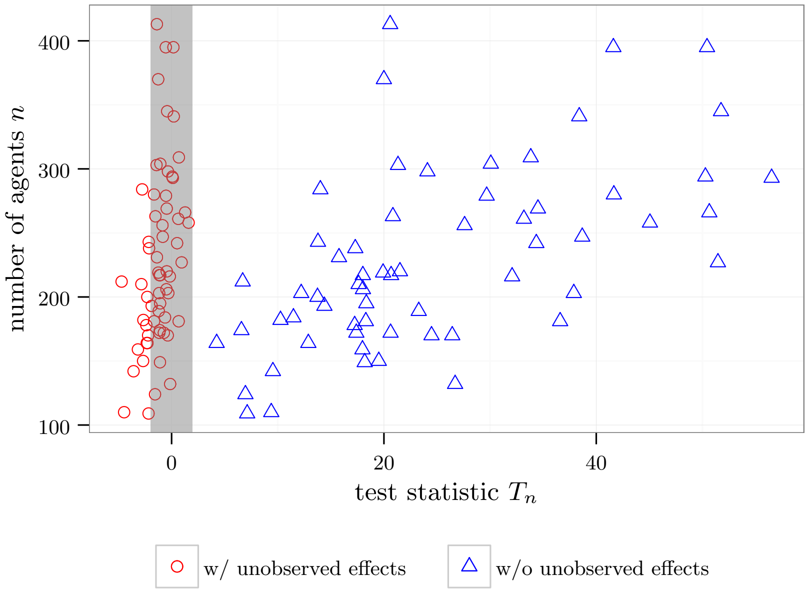


Figure 3.5: Comparing test statistics under the model with unobserved effects to test statistics under the model without unobserved effects for all networks in data set. The shaded region gives the interval in which a two-sided test does not reject at level $\alpha = .05$.

the null is rejected only for about a quarter of the villages. It is unavoidable that passing to a more complex model adds additional statistical noise. However, the differences in test results are only partially due to larger standard errors in the model with unobserved effects.

All T_n-values for the parametric model are positive, suggesting that the linking model underestimates the network's tendency towards transitive closure. A popular candidate for the cause of such a failure is the notion that agents derive utility from transitive relations. If this were true, then agents would care about endogenous attributes of the network, violating the exogeneity assumption. For the model with unobserved effects the test statistic takes on positive as well as negative values. All rejections are for negative values of T_n. While this is still suggestive of non-random behavior, it invites a fundamentally different interpretation of the way in which the model fails. A distaste for transitivity does not have much theoretical appeal, leading the researcher to consider other mechanisms, such as systematic under-reporting of transitive relations in the survey.

This illustrates well that specification tests can offer more than a binary indication of model validity. Some rejections provide evidence that the model misrepresents the economic context in a fundamental way. Other rejections have less severe repercussions and might still allow the researcher to maintain the model as a useful approximation.

For the favor networks in Indian networks, it seems that accounting for unobserved sources of degree heterogeneity may be sufficient to dismantle circumstantial evidence for network externalities.

3.7 Conclusion

The ideas explored in this paper open up several avenues for future research into network formation models.

An obvious extension is to replace independence of the link-specific shocks by a less restrictive exogeneity assumption. It is natural to allow for correlation between the two shocks that are relevant for the links between a given pair of agents. This can be accomplished by passing to a model that imposes an *iid* assumption on tuples $(\epsilon_{(i,j)}, \epsilon_{(j,i)})$, $i \neq j$. Such a model is a network version of a bivariate probit model with fixed effects. The analysis of this model can proceed along similar steps as those outlined in this paper for the simpler model. In the bivariate model, the correlation between the within-dyad shocks is an additional parameter of economic interest. Similar to a corresponding parameter in the model of Holland and Leinhardt 1981 this correlation describes agent preferences for reciprocating links. An alternative approach is to put more structure on the dyadic interaction by formulating link formation as an appropriate multinomial choice problem that lets each pair of agents jointly decide which of the four possible link configurations within the dyad they want to have. This approach requires that the economist has sufficient prior knowledge about the nature of the dyadic interaction to set up a meaningful choice model.

I have presented results for transitivity as an example of local structure. Depending on the specific application in mind other features might be of interest. It is an interesting challenge to provide a unified theory of inference in the presence of unobserved heterogeneity for a broad class of local network features. The difficulty of such an endeavor lies in finding a general expression for the asymptotic bias.

In this paper, I focus on a relatively simple transitivity measure. This is mainly for expositional convenience. In fact, the asymptotic distribution of a plug-in estimator of the clustering coefficient from equation (3.4) is provided by a straightforward corollary to my results. To see this, note that upon suitable normalization of the clustering coefficient, the denominator can be replaced by its probability limit at the expense of an $o_p(1)$ term. Then, Theorem 3.3 and an appeal to the delta-method give the desired distribution.

My transitivity test improves on previous tests (Holland and Leinhardt 1978; Karlberg 1999) along two dimensions. First, it is asymptotically normal. Approximate critical values can be obtained from the asymptotic distribution. This obviates the need for computationally intensive simulations. Secondly, it explicitly takes into account the estimation error from estimating the reference distribution. This caters to many empirical applications in which knowledge about an appropriate reference distribution is limited. These two achievements are possible because of the way the model controls for degree heterogeneity. The unobserved effects approach allows for a flexible degree distribution while also admitting a large sample theory. It seems that other testing problems in networks could benefit from this framework as well. Further research is needed.

As I have shown above, controlling for unobserved heterogeneity is essential for giving an accurate description of local network features. In my application, a model that does not admit unobserved heterogeneity estimates spurious excess transitivity. Controlling for unobserved sources of degree heterogeneity *reverses* the verdict regarding transitivity. The level of transitivity predicted from the model without unobserved effects is not significantly higher than the observed level of transitivity. Recently, the econometric research on network formation models has focused on allowing agents to care for endogenous network attributes. Some significant progress has been made in this direction (Mele 2013; Sheng 2014; Miyauchi 2014; Leung 2014), but this has come at the expense of neglecting ramifications of unobserved heterogeneity. In particular, as I argue for transitivity, unobserved heterogeneity and agent preferences for endogenous network features ("network externalities") are competing explanations when it comes to justifying the prevalence of certain local structures. In some economic settings one explanation might seem more plausible, for others, the other explanation is more compelling. In settings in which both explanations have a claim to validity, identification strategies will have to be developed to disentangle the two effects.

Appendix 3.A Notation

In this part of the appendix I introduce notation from Fernández-Val and Weidner 2014 (henceforth FVW) that will be helpful in the subsequent proofs. We let $\phi = (\gamma_1^S, \dots, \gamma_n^S, \gamma_1^R, \dots, \gamma_n^R)$ denote the incidental parameter vector. The unobserved effect for the link (i,j) is $\pi_{(i,j)} = \gamma_i^S + \gamma_j^R$. The likelihood contribution of edge e is

$$\begin{aligned}\ell_e(\theta, \phi) &= Y_e \log p_e + (1 - Y_e)\log(1 - p_e)\\ &= Y_e \log F_e(X_e\theta + \pi_e) + (1 - Y_e)\log(1 - F_e(X_e\theta + \pi_e)).\end{aligned}$$

We write $\ell_e = \ell_e(\theta^0, \phi^0)$ for the likelihood contribution evaluated at the true parameters. Note that $\partial_\pi \ell_e = H_e(Y_e - p_e)$ and $\partial_\theta \ell_e = (\partial_\pi \ell_e)X_e$ (compare also Example 1 in FVW). The empirical likelihood is

$$\mathcal{L}(\theta, \phi) = \frac{1}{n}\sum_e \ell_e(\theta, \phi) - b\left((\iota_n', -\iota_n')\phi\right)^2/2,$$

where the last term is a penalty that imposes the restriction $\sum_i(\gamma_i^S - \gamma_i^R) = 0$ on the incidental parameter and b is an arbitrary positive constant. We write $\mathcal{L} = \mathcal{L}(\theta, \phi)$ and $\bar{\mathcal{L}} = \bar{\mathbb{E}}\,\mathcal{L}$ and use corresponding notation for the derivatives of the likelihood. Furthermore we let

$$\mathcal{S} = \partial_\phi \mathcal{L} = \begin{pmatrix} \left[\frac{1}{n}\sum_{j:j\neq i} \partial_\pi \ell_{(i,j)}\right]_{i\in V} \\ \left[\frac{1}{n}\sum_{i:i\neq j} \partial_\pi \ell_{(i,j)}\right]_{j\in V} \end{pmatrix}$$

denote the likelihood score with respect to the incidental parameter evaluated at the true parameters and let $\mathcal{H} = -\partial_{\phi\phi'}\mathcal{L}$ denote the corresponding Hessian. Let $\bar{\mathcal{H}} = \bar{\mathbb{E}}\,\mathcal{H}$ and $\tilde{\mathcal{H}} = \mathcal{H} - \bar{\mathcal{H}}$.

Appendix 3.B Proofs of main theorems

Proof of Theorem 3.1 The proof follows the same line of arguments as the proof of Theorem 4.1 in FVW. The only difference is that, while in the panel set-up all time periods are observed for each individual, in the network model only $n-1$ out of n possible links to receivers are permitted. Namely, my model does not allow self-links. This, however, can be accommodated. For example, note that the score with respect to the incidental parameter can be written as

$$\mathcal{S} = \begin{bmatrix} M\iota_n \\ M'\iota_n \end{bmatrix}$$

where M is the $n \times n$ matrix with entries $M_{i,j} = \partial_\pi \ell_{(i,j)}$ for $i \neq j$ and $M_{i,j} = 0$ otherwise. This is the representation assumed in the application of Lemma D.11 of FVW and one can proceed as in their proof. For the other projection arguments one proceeds similarly. □

Proof of Theorem 3.2 The theorem follows form an expansion in the proof of Theorem 3.5. □

Proof of Theorem 3.3 The theorem follows from an expansion in the proof of Theorem 3.6. □

Proof of Theorem 3.4 The result follows from Theorem 3.6 in conjunction with Corollary 3.1 setting $a_n = a = 1$. □

Proof of Theorem 3.5 Let $\hat{\rho}_\beta = \prod_{e\in\beta} p_e(X_e, \hat{\pi}_e, \hat{\beta})$. We decompose

$$n^{-2}\left(S_n - \widehat{\mathrm{E}\, S_n}\right) = n^{-2}\sum\nolimits_\beta (T_e - \rho_\beta) - \left(n^{-2}\sum\nolimits_\beta \hat{\rho}_\beta - n^{-2}\sum\nolimits_\beta \rho_\beta\right).$$

For the first term, argue similarly to the proof of Theorem 3.6 that

$$n^{-2}\sum\nolimits_\beta (T_e - \rho_\beta) = n^{-1}\sum_e (Y_e - p_e)\,\boldsymbol{\beta}_e^n + o_p(1).$$

For the second term let $\Delta = n^{-3}\sum_\beta \rho_\beta$ and $\hat{\Delta} = n^{-3}\sum_\beta \hat{\rho}_\beta$. Note that Δ behaves like the partial effect considered in FVW and employ their Theorem B.4 to show that conditional on observables and unobserved effects

$$\hat{\Delta} - \Delta = \left[\partial_{\theta'}\Delta + (\partial_{\phi'}\Delta)\bar{\mathcal{H}}^{-1}(\partial_{\phi\theta'}\bar{\mathcal{L}})\right](\hat{\theta} - \theta^0) + U_\Delta^{(0)} + U_\Delta^{(1)} + o_p\left(n^{-1}\right)$$

with

$$\begin{aligned} U_\Delta^{(0)} =& (\partial_{\phi'}\Delta)\bar{\mathcal{H}}^{-1}\mathcal{S}, \\ U_\Delta^{(1)} =& -(\partial_{\phi'}\bar{\Delta})\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\mathcal{S} \\ &+ \frac{1}{2}\mathcal{S}'\bar{\mathcal{H}}^{-1}\left[\partial_{\phi\phi'}\Delta + \sum_{g=1}^{\dim\phi}\left[\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}}\right]\left[\bar{\mathcal{H}}^{-1}(\partial_{\phi'}\Delta)\right]_g\right]\bar{\mathcal{H}}^{-1}\mathcal{S}. \end{aligned}$$

Define the $n \times n$ matrix D^n by

$$D^n = \begin{cases} D^n_{(i,j)} & \text{for } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

and note that

$$\partial_\phi \Delta = \frac{1}{n^2} \begin{pmatrix} D^n \iota_N \\ (D^n)' \iota_N \end{pmatrix}$$

and

$$\partial_\theta \Delta = \frac{1}{n^2} \sum_{e \in E} \left\{ \frac{1}{3n} \sum_{\beta \ni e} \partial_\theta \rho_\beta \right\} = \frac{1}{n^2} \sum_{e \in E} \left\{ (\partial_\theta p_\beta) \frac{1}{n} \sum_{\beta \ni e} \rho_{-e}(\beta) \right\} = \frac{1}{n^2} \sum_{e \in E} (\partial_\pi p_\beta) \boldsymbol{\beta}^n_e X_e .$$

Using $\partial_\pi p_\beta = f_e$ the projection argument from Lemma B.11 in FVW gives

$$\partial_{\theta'} \Delta + (\partial_{\phi'} \Delta) \bar{\mathcal{H}}^{-1} (\partial_{\phi \theta'} \bar{\mathcal{L}}) = \Xi_n .$$

Arguments similar to the ones employed in the proofs of Theorem C.1 and Theorem 4.2 in FVW give

$$\begin{aligned} & n \left(\frac{1}{2} \mathcal{S}' \bar{\mathcal{H}}^{-1} \left[\sum_{g=1}^{\dim \phi} \left[\partial_{\phi \phi' \phi_g} \bar{\mathcal{L}} \right] \left[\bar{\mathcal{H}}^{-1} (\partial_{\phi'} \Delta) \right]_g \right] \bar{\mathcal{H}}^{-1} \mathcal{S} - (\partial_{\phi'} \Delta) \bar{\mathcal{H}}^{-1} \tilde{\mathcal{H}} \bar{\mathcal{H}}^{-1} \mathcal{S} \right) \\ &= \lim_{n \to \infty} \frac{1}{2n} \sum_i \frac{\sum_{j:j \neq i} H_{(i,j)} \Psi_{(i,j)} \partial f_{(i,j)}}{\sum_{j:j \neq i} \omega_{(i,j)}} + \lim_{n \to \infty} \frac{1}{2n} \sum_j \frac{\sum_{i:i \neq j} H_{(i,j)} \Psi_{(i,j)} \partial f_{(i,j)}}{\sum_{i:i \neq j} \omega_{(i,j)}} + o_p(1). \end{aligned}$$

For the remaining term the arguments in FVW do not apply as $\partial_{\phi \phi'} \Delta$ does not exhibit as symmetric a structure as the corresponding derivative of a partial effect. Instead write

$$n \left(\partial_{\phi \phi'} \Delta \right) = \begin{bmatrix} A^\phi_{SS} & A^\phi_{SR} \\ \left(A^\phi_{SR} \right)' & A^\phi_{RR} \end{bmatrix} = \begin{bmatrix} \bar{A}^\phi_{SS} + \tilde{A}^\phi_{SS} & \bar{A}^\phi_{SR} + \tilde{A}^\phi_{SR} \\ \left(\bar{A}^\phi_{SR} + \tilde{A}^\phi_{SR} \right)' & \bar{A}^\phi_{SS} + \tilde{A}^\phi_{RR} \end{bmatrix}$$

with $\bar{A}^\phi_k$ a diagonal matrix such that $\| \bar{A}^\phi_k \|_{\max} = O_p(1)$ and $\tilde{A}^\phi_k$ such that $\| \bar{A}^\phi_k \|_{\max} = O_p(n^{-1})$ for $k \in \{SS, SR, RR\}$. By Lemma D.8 in FVW the expected Hessian with respect to the incidental parameter has the same structure

$$\bar{\mathcal{H}}^{-1} = \begin{bmatrix} \bar{\mathcal{H}}^{-1}_{SS} & \bar{\mathcal{H}}^{-1}_{SR} \\ (\bar{\mathcal{H}}^{-1}_{SR})' & \bar{\mathcal{H}}^{-1}_{RR} \end{bmatrix} = \begin{bmatrix} \bar{\bar{\mathcal{H}}}^{-1}_{SS} + \tilde{\bar{\mathcal{H}}}^{-1}_{SS} & \bar{\bar{\mathcal{H}}}^{-1}_{SR} + \tilde{\bar{\mathcal{H}}}^{-1}_{SR} \\ \left(\bar{\bar{\mathcal{H}}}^{-1}_{SR} + \tilde{\bar{\mathcal{H}}}^{-1}_{SR} \right)' & \bar{\bar{\mathcal{H}}}^{-1}_{RR} + \tilde{\bar{\mathcal{H}}}^{-1}_{RR} \end{bmatrix} .$$

Now note that, for D_1, D_2 diagonal stochastic matrices with $\|D_k\|_{\max} = O_p(1)$, $k = 1, 2$, and M_1, M_2 stochastic matrices such that $\|M_k\|_{\max} = O_p(n^{-1})$, $k = 1, 2$, $D_1 \times D_2$ is a stochastically bounded diagonal matrix, and $D_1 \times M_1$ and $M_1 \times M_2$ are bounded by an $O_p(n^{-1})$ term. All bounds are in terms of the matrix maximum norm. Let Υ denote a

$n \times n$ random matrix with entries $\Upsilon_{i,j} = \partial_\pi \ell_{(i,j)}$ if $i \neq j$ and $\Upsilon_{i,j} = 0$ otherwise. The score with respect to the incidental parameter can be written as

$$\mathcal{S} = \begin{bmatrix} \mathcal{S}_S \\ \mathcal{S}_R \end{bmatrix} = \frac{1}{n} \begin{bmatrix} \Upsilon \, \iota_n \\ \Upsilon' \, \iota_n \end{bmatrix}.$$

Multiplying out the partitioned matrices and employing Lemma 3.2 multiple times gives

$$n \left(\mathcal{S}' \bar{\mathcal{H}}^{-1} \partial_{\phi\phi'} \Delta \bar{\mathcal{H}}^{-1} \mathcal{S} \right) = \mathrm{E} \left[\mathcal{S}_S' \bar{\bar{\mathcal{H}}}_{SS}^{-1} \bar{A}_{SS}^{\phi} \bar{\bar{\mathcal{H}}}_{SS}^{-1} \mathcal{S}_S \right] + \mathrm{E} \left[\mathcal{S}_R' \bar{\bar{\mathcal{H}}}_{RR}^{-1} \bar{A}_{RR}^{\phi} \bar{\bar{\mathcal{H}}}_{RR}^{-1} \mathcal{S}_R \right] + o_p(1).$$

Now

$$\begin{aligned} \mathrm{E} \left[\mathcal{S}_S' \bar{\bar{\mathcal{H}}}_{SS}^{-1} \bar{A}_{SS}^{\phi} \bar{\bar{\mathcal{H}}}_{SS}^{-1} \mathcal{S}_S \right] &= \frac{1}{n^2} \mathrm{E} \left\{ \sum_i (\bar{A}_{SS}^{\phi})_{i,i} \frac{\sum_{j:j \neq i} \left(\partial_\pi \ell_{(i,j)} \right)^2}{\left(\frac{1}{n} \sum_{j:j \neq i} \omega_{(i,j)} \right)^2} \right\} \\ &= \frac{1}{n} \sum_i \frac{(n-1)(\bar{A}_{SS}^{\phi})_{i,i}}{\sum_{j:j \neq i} \omega_{(i,j)}} + o(1), \end{aligned}$$

where the second equality follows from a Bartlett identity. By symmetry

$$\mathrm{E} \left[\mathcal{S}_R' \bar{\bar{\mathcal{H}}}_{RR}^{-1} \bar{A}_{RR}^{\phi} \bar{\bar{\mathcal{H}}}_{RR}^{-1} \mathcal{S}_R \right] = \frac{1}{n} \sum_j \frac{(n-1)(\bar{A}_{RR}^{\phi})_{j,j}}{\sum_{i:i \neq j} \omega_{(i,j)}} + o(1).$$

Since $\left(\bar{A}_{SS}^{\phi} \right)_{i,i} = \delta_i^S$ and $\left(\bar{A}_{RR}^{\phi} \right)_{j,j} = \delta_j^R$,

$$\begin{aligned} n \left(U_\Delta^{(1)} \right) = & \lim_{n \to \infty} \frac{1}{2n} \sum_i \frac{\sum_{j:j \neq i} \left(H_{(i,j)} \Psi_{(i,j)} \partial f_{(i,j)} + \delta_i^S \right)}{\sum_{j:j \neq i} \omega_{(i,j)}} \\ & + \lim_{n \to \infty} \frac{1}{2n} \sum_j \frac{\sum_{i:i \neq j} \left(H_{(i,j)} \Psi_{(i,j)} \partial f_{(i,j)} + \delta_j^R \right)}{\sum_{i:i \neq j} \omega_{(i,j)}} + o_p(1). \end{aligned}$$

Using similar arguments as in the proofs of Theorem C.1 in FVW one can show that

$$U_\Delta^{(0)} = -\frac{1}{n} \sum_e \Psi_e \partial_\pi \ell_e.$$

From the proof of Theorem 3.3

$$\tilde{W}_\infty \, n \left(\hat{\theta} - \theta^0 \right) = B_\infty + D_\infty + \frac{1}{n} \sum_e \left(\partial_\beta \ell_e - \partial_\pi \ell_e (X_e - \tilde{X}_e) \right) + o_p(1).$$

Plugging in for the binary choice model gives

$$\partial_\pi \ell_e = H_e (Y_e - p_e) \quad \text{and} \quad \partial_\beta \ell_e = (\partial_\pi \ell_e) \, X_e.$$

Therefore, the stochastic part of $n^{-2}\left(S_n - \widehat{\mathrm{E}\, S_n}\right)$ can be written as

$$\begin{aligned}
&\frac{1}{n}\sum_e (Y_e - p_e)\left(\boldsymbol{\beta}_e^n - H_e(\Xi_n' \bar{W}_\infty^{-1}\tilde{X}_e - \Psi_e)\right)\\
=&\frac{1}{n}\sum_e (Y_e - p_e)\left(\boldsymbol{\beta}_e^n - H_e(\Xi_\infty' \bar{W}_\infty^{-1}\tilde{X}_e - \Psi_e)\right)\\
&-\frac{1}{n}\sum_e (Y_e - p_e) H_e\left((\Xi_n - \Xi_\infty)' \tilde{W}_\infty^{-1}\tilde{X}_e - \Psi_e\right).
\end{aligned}$$

It can be shown by standard arguments that the second term is $o_p(1)$. For the first term, an appeal to the Lindeberg-Feller central limit theorem gives the desired normal distribution. Collecting terms gives an asymptotic bias of $B_\infty^{TT} + D_\infty^{TT}$. □

Appendix 3.C Auxiliary results

3.C.1 Example 3.1

The claim in the example follows from the following lemma.

Lemma 3.1 *Let $\rho_n^\circ = \frac{n_n^\circ}{n} p_n^\circ$ and let $\rho_n^\star = \frac{n_n^\star}{n} p_n^\star$. Then*

$$\liminf_n \frac{\mathrm{E}_{M_{simple,n}}\left[\#\ transitive\ triples\right]}{\mathrm{E}_{M_{simple,n}^{proj}}\left[\#\ transitive\ triples\right]} \geq 1 + \liminf_n e_n.$$

with

$$e_n = \frac{(p_n^\star - p_n^\circ)^2}{p_n^\star p_n^\circ}\left(\frac{\rho_n^\star}{\rho_n^\circ}\right)\left(1 + \frac{\rho_n^\star}{\rho_n^\circ}\right)^{-2}.$$

Proof Let R_n denote the ratio on the right-hand side. For a positive integer m define the factorial power $m^{\underline{k}} = m(m-1)\cdots(m-k+1)$. We first ignore the n subscript and the asymptotic framework and give an exact calculation for fixed n. For the denominator of the ratio above we can write

$$\mathrm{E}_{M_{\text{simple}}^{\text{proj}}}\left[\#\ \text{transitive triples}\right] = n^{\underline{3}} p^3 = n^{\underline{3}}\left(\frac{n^\circ}{n}p^\circ + \frac{n^\star}{n}p^\star\right) \overset{def}{=} n^{\underline{3}} D_n.$$

Turning to the nominator we partition all transitive triples (TTs) by the number of attractor nodes that they contain.

$\boxed{\text{0 attractors}}$ The number of TTs with exactly zero attractor nodes is

$$\binom{n^\circ}{3}\mathrm{Iso}(\text{transitive triple}) = (n^\circ)^{\underline{3}},$$

where Iso(G) is the number of isomorphisms of the graph G. Since all positions in a transitive triple are unique, the number of isomorphisms of a transitive triple is equal to the permutations of node labels, i.e., Iso(transitive triple) = 3!. Each of these TTs has probability $(p^\circ)^3$. The contribution to the expectation is $(n^\circ)^{\underline{3}}(p^\circ)^3$.

$\boxed{\text{1 attractor}}$ There are $\binom{n^\circ}{2}n^\star$ ways of selecting the nodes. Given three nodes there are

2! TTs with probability $(p^\circ)^3$

2! TTs with probability $(p^\circ)^2 p^\star$

2! TTs with probability $p^\circ (p^\star)^2$.

In sum, the contribution of these TTs to the expectation is

$$n^\star (n^\circ)^{\underline{2}} (p^\circ)^2 p^\star \left(1 + \frac{p^\circ}{p^\star} + \frac{p^\star}{p^\circ}\right) = (n^\circ)^{\underline{2}} n^\star (p^\circ)^2 p^\star (3 + w_n),$$

where

$$w_n = \frac{(p^\star - p^\circ)^2}{p^\star p^\circ}.$$

$\boxed{\text{2 attractors}}$ There are $n^\circ \binom{n^\star}{2}$ ways of selecting the nodes. Given three nodes there are

2! TTs with probability $(p^\star)^3$

2! TTs with probability $(p^\star)^2 p^\circ$

2! TTs with probability $p^\star (p^\circ)^2$.

The contribution of these TTs to the expectation is

$$n^\circ \binom{n^\star}{2} (p^\star)^2 p^\circ \left(1 + \frac{p^\circ}{p^\star} + \frac{p^\star}{p^\circ}\right) = n^\circ (n^\star)^{\underline{2}} p^\circ (p^\star)^2 (3 + w_n),$$

where w_n is defined as above.

$\boxed{\text{3 attractors}}$ Arguing as above it is easy to see that the contribution to the expectation is $(n^\star)^{\underline{3}} (p^\star)^3$.

Putting the results from above together we get

$$\begin{aligned}&\mathrm{E}_{M_{\text{simple}}}\left[\#\text{ transitive triples}\right]\\ =&(n^\circ)^{\underline{3}} (p^\circ)^3 + (3 + w)(n^\circ)^{\underline{2}} n^\star (p^\circ)^2 p^\star + (3 + w) n^\circ (n^\star)^{\underline{2}} p^\circ (p^\star)^2 + (n^\star)^{\underline{3}} (p^\star)^3.\end{aligned}$$

Returning to the asymptotic framework, dividing nominator and denominator by $n^{\underline{3}}$ and expanding D_n it is now easy to see that

$$\begin{aligned}R_n &= 1 + w_n \frac{\frac{(n^\circ)^{\underline{2}} n^\star}{n^{\underline{3}}} (p^\circ)^2 p^\star + \frac{n^\circ (n^\star)^{\underline{2}}}{n^{\underline{3}}} p^\circ (p^\star)^2}{D_n} + o(1)\\ &= 1 + w_n \frac{(\rho_n^\circ)^2 \rho_n^\star + \rho_n^\circ (\rho^\star)^2}{D_n} + o(1).\end{aligned}$$

Since

$$D_n = (\rho_n^\circ)^3 + 3(\rho_n^\circ)^2 \rho^\star + 3\rho_n^\circ (\rho_n^\star)^2 + (\rho_n^\star)^3$$

we have

$$\frac{D_n}{(\rho_n^\circ)^2 \rho_n^\star} = f\left(\frac{\rho_n^\star}{\rho_n^\circ}\right).$$

for $f(x) = 3 + x^2 + 3x + x^{-1}$ and hence by symmetry

$$\frac{D_n}{\rho_n^\circ (\rho^\star)^2} = f\left(\frac{\rho_n^\circ}{\rho_n^\star}\right).$$

Now noting that $f(x^{-1}) = xf(x)$ straightforward calculations yield

$$[f(x)]^{-1} + \left[f\left(x^{-1}\right)\right]^{-1} = \frac{x}{(1+x)^2}.$$

□

3.C.2 Sparse dyadic model without unobserved effects

In this appendix we consider a variation of the model (3.3) where the link function is allowed to depend on n. We assume that the link function is given by $F_n = a_n^{-1} F$ for a deterministic sequence a_n and a base link function F. For $a_n \to \infty$ this allows for asymptotically sparse networks. Both a_n and F are assumed to be known. For notational convenience we redefine $\rho_{-e}(\beta) = \prod_{\substack{\beta_j \in \beta \\ \beta_j \neq e}} F_e$.

Theorem 3.6 *Suppose that $a_n \geq 1$ and $a_n^{-1} n^2 \to \infty$. Assume that the base link function F is bounded away from zero and one, and that it is three times continuously differentiable. Let $R_{\theta,\infty} = \lim_{n\to\infty} R_{\theta,n}$,*

$$\breve{W}_\infty = \lim_{n\to\infty} \frac{1}{n^2} \sum_{e \in E(n)} \frac{f_e^2}{F_e(1 - a_n^{-1} F_e)} \mathcal{X}_e \mathcal{X}_e',$$

$$\breve{V}_S^{(a)} = \lim_{n\to\infty} \frac{1}{n^2} \sum_{e \in E(n)} F_e(1 - a_n^{-1} F_e) \left\{\boldsymbol{\beta}_e^n - X_e^\Diamond\right\}^2,$$

$$\breve{V}_S^{(b)} = \lim_{n\to\infty} \frac{1}{n^2} \sum_{e \in E(n)} \sum_{\substack{\beta,\beta' \\ \beta \cap \beta' = \{e\} \\ |V(\beta) \cap V(\beta)| = 2}} \frac{(\rho_{-e}(\beta) - \frac{1}{3} X_e^\Diamond)(\rho_{-e}(\beta') - \frac{1}{3} X_e^\Diamond)}{n^2} F_e(1 - a_n^{-1} F_e),$$

where $X_e^\Diamond = \left(R_{\theta,\infty}\right)' \breve{W}_\infty^{-1} f_e \mathcal{X}_e / (F_e(1 - a_n^{-1} F_e))$. Suppose that conditional on $\mathbf{X}$ all limits exist and that $\breve{V}_S^{(a)} > 0$. Then conditionally on $\mathbf{X}$

$$n^{-2}\left(S_n - \widehat{\mathrm{E}\, S_n}\right) \xrightarrow{d} \mathcal{N}\left(0, \breve{V}_S^{(a)}\right).$$

Moreover,

$$\frac{\breve{V}_S^{(b)}}{\breve{V}_S^{(a)}} \to 1.$$

Proof We work conditionally on $\mathbf{X}$. First, we compute the variance of S_n. Since triangles β and β' are independent provided that $\beta \cap \beta' = \varnothing$ we get

$$\begin{aligned}\operatorname{var} S_n =& \mathrm{E}\left(\sum_{\beta\in B}(T_\beta - \mathrm{E}\, T_\beta)\right)^2\\ =& a_n^{-4} \sum_{\substack{(\beta,\beta')\in B\times B\\ |\beta\cap\beta'|=1}} \mathrm{E}\left[(T_\beta - ET_\beta)(T_{\beta'} - ET_{\beta'})\right] + H_n\\ =& a_n^{-5} \sum_e \sum_{\substack{(\beta,\beta')\in B\times B\\ \beta\cap\beta'=\{e\}}} F_e(1 - a_n^{-1}F_e)\rho^t_{-e}(\beta)\rho^t_{-e}(\beta') + H_n,\end{aligned}$$

where H_n captures the contribution to the expectation from triangle pairs that share 2 or 3 edges. The number of triangle pairs that share 2 edges and the number of triangle pairs that share 3 edges (these are just the pairs (β,β), $\beta \in B$) are both of the same order as n^3. Note that since F is bounded away from zero there is a constant C_1 such that the contribution to the expectation is less than $C_1 a_n^{-4}$ and $C_1 a_n^{-3}$ for each pair of triangles with 2 and 3 common nodes, respectively. Hence,

$$R_n = O\left(a_n^{-4}n^3 + a_n^{-3}n^3\right) = O\left(a_n^{-3}n^3\right).$$

Let $\hat{S}_n$ denote the Hajek-Projection of $S_n - \mathrm{E}\, S_n$ onto the $(Y_e)_{e\in E}$, i.e.,

$$\hat{S}_n = \sum_e \mathrm{E}\left[S_n - \mathrm{E}\, S_n \mid Y_e\right] = \sum_e \sum_\beta \mathrm{E}\left[T_\beta - \mathrm{E}\, T_\beta \mid Y_e\right].$$

Obviously, $\mathrm{E}\left[T_\beta - \mathrm{E}\, T_\beta \mid Y_e\right] = 0$ if $e \notin \beta$. Otherwise,

$$\mathrm{E}\left[T_\beta - \mathrm{E}\, T_\beta \mid Y_e\right] = a_n^{-2}\left(Y_e - a_n^{-1}F_e\right)\rho_{-e}(\beta).$$

Therefore,

$$\hat{S}_n = a_n^{-2} \sum_e \sum_{\beta\ni e} \left(Y_e - a_n^{-1}F_e\right)\rho_{-e}(\beta)$$

and

$$\begin{aligned}\operatorname{var} \hat{S}_n &= a_n^{-4} \sum_e \sum_{\substack{(\beta,\beta')\in B\times B\\ \beta\cap\beta'=\{e\}}} \rho_{-e}(\beta)\rho_{-e}(\beta')\,\mathrm{E}\left(Y_e - a_n^{-1}F_e\right)^2\\ &= a_n^{-5} \sum_e \sum_{\substack{(\beta,\beta')\in B\times B\\ \beta\cap\beta'=\{e\}}} F_e(1 - a_n^{-1}F_e)\rho_{-e}(\beta)\rho_{-e}(\beta').\end{aligned}$$

As F is bounded away from zero for some constant C

$$\operatorname{var} \hat{S}_n \geq C a_n^{-5} n^4$$

and therefore

$$\frac{\operatorname{var} S_n}{\operatorname{var} \hat{S}_n} \leq 1 + O\left(\frac{1}{a_n^{-2} n}\right).$$

As $H_n \geq 0$ we also have $\operatorname{var} S_n \geq \hat{S}_n$ and therefore

$$\frac{\operatorname{var} S_n}{\operatorname{var} \hat{S}_n} \to 1$$

so that by Theorem 11.2 in van der Vaart 2000

$$S_n - \mathrm{E}\, S_n - \hat{S}_n = o_p\left(\sqrt{\operatorname{var} \hat{S}_n}\right) = o_p\left(a_n^{5/2} n^{-2}\right).$$

Then

$$\begin{aligned} S_n - \widehat{\mathrm{E}\, S_n} &= S_n - \mathrm{E}\, S_n - \left(\widehat{\mathrm{E}\, S_n} - \mathrm{E}\, S_n\right) \\ &= \hat{S}_n - \left(\widehat{\mathrm{E}\, S_n} - \mathrm{E}\, S_n\right) + o_p\left(a_n^{5/2} n^{-2}\right). \end{aligned}$$

Turning first to the second term note that

$$\begin{aligned} \widehat{\mathrm{E}\, S_n} - \mathrm{E}\, S_n =& a_n^{-3} n^3 R_{\theta,n}(\theta^0)\big(\hat{\theta} - \theta^0\big) \\ &+ \frac{1}{2} a_n^{-3} \big(\hat{\theta} - \theta^0\big)' \Big[\sum_{\beta} \partial_{\theta\theta'} (F_{\beta_1} F_{\beta_2} F_{\beta_3})(\tilde{\theta})\Big] \big(\hat{\theta} - \theta^0\big) \\ =& a_n^{-3} n^3 R_{\theta,n} \big(\hat{\theta} - \theta^0\big) + O_p\left(a_n^{-3} n^3 \|\hat{\theta} - \theta^0\|^2\right), \end{aligned}$$

where $\tilde{\theta}$ is an intermediate value. It is easy to show that $\hat{\theta}$ has an asymptotically linear representation

$$a_n^{-1} n \left(\hat{\theta} - \theta^0\right) = \frac{1}{n} \sum_{e \in E(n)} \psi_e + O_p\left(\frac{1}{a_n^{-1} n}\right)$$

with influence function

$$\psi_e = \breve{W}_\infty^{-1} \frac{f_e}{F_e(1 - a_n^{-1} F_e)} \mathcal{X}_e(Y_e - a_n^{-1} F_e).$$

Plugging in the linear representation gives

$$\begin{aligned} a_n^{5/2} \frac{\widehat{\mathrm{E}\, S_n} - \mathrm{E}\, S_n}{n^2} =& a_n^{1/2} R_{\theta,n} a_n^{-1} n \big(\hat{\theta} - \theta^0\big) + o_p(1) \\ =& \frac{a_n^{1/2}}{n} \sum_{e \in E(n)} X_e^{\lozenge} \left(Y_e - a_n^{-1} F_e\right) + o_p(1). \end{aligned}$$

Therefore

$$a_n^{5/2} \frac{S_n - \widehat{\mathrm{E}\, S_n}}{n^2} = \frac{a_n^{1/2}}{n} \sum_{e \in E(n)} \left(\frac{\sum_{\beta \ni e} \rho_{-e}(\beta)}{n} - X_e^{\lozenge}\right) \left(Y_e - a_n^{-1} F_e\right) + o_p(1).$$

The variance of the first term on the right-hand side is $\breve{V}_S^{(a)} + o(1)$. It is straightforward to verify that Lindeberg's condition is satisfied. The claim about $\breve{V}_S^{(b)}$ follows by noting that

$$\left(\frac{\sum_{\beta \ni e} \rho_{-e}(\beta)}{n} - X_e^{\lozenge}\right)^2 = \left(\sum_{\beta \ni e} \frac{\rho_{-e}(\beta) - \frac{1}{3} X_e^{\lozenge}}{n}\right)^2$$

and expanding the square. It is then easy to see that the resulting sum is dominated by pairs of triangles that share exactly two vertices. □

Under a distributional assumption about the $\mathcal{X}_e$ the limits in Theorem 3.6 can be shown to exist and the asymptotic variance can be expressed as a function of subgraphs on the vertex set $\{1,2,3,4\}$. To this end, let

$$B_{+v} = \{\beta : \beta \text{ is a TT on } \{1,2,v\}, (1,2) \in \beta\}.$$

Corollary 3.1 *Suppose that the assumptions of Theorem 3.6 and in addition Assumption 3.2 hold. Suppose also that $a_n \to a$. Let*

$$V_\theta^{(c)} = \sum_{\substack{\beta \in B_{+3} \\ \beta' \in B_{+4}}} \mathrm{E}\Big\{\Big(\rho_{-(1,2)}(\beta) - \frac{1}{3}X^\lozenge_{(1,2)}\Big)\Big(\rho_{-(1,2)}(\beta') - \frac{1}{3}X^\lozenge_{(1,2)}\Big)F_{(1,2)}\left(1 - a^{-1}F_{(1,2)}\right)\Big\}.$$

Then

$$\breve{W}_\infty = \mathrm{E}\Big\{\frac{f_e^2}{F_e(1-a^{-1}F_e)}\mathcal{X}_e\mathcal{X}_e'\Big\},$$
$$R_{\theta,\infty} = \mathrm{E}\left\{\partial_\theta \rho_\beta(\mathbf{X}, \theta^0)\right\}$$

on a set with probability approaching one and

$$n^{-2}a_n^{5/2}\big(S_n - \widehat{\mathrm{E}\,S_n}\big) \xrightarrow{d} \mathcal{N}\big(0, \breve{V}_S^{(c)}\big).$$

Proof The first two statements follow by standard arguments using the Markov inequality. Note that $\breve{V}_S^{(c)} = \mathrm{E}\,\breve{V}_S^{(b)}$. The distributional result follows by Theorem 3.6 if we show

$$\frac{\breve{V}_S^{(b)}}{\mathrm{E}\,\breve{V}_S^{(b)}} \xrightarrow{p} 1.$$

It suffices to show that the variance of the ratio on the left-hand side vanishes. To this end, let

$$\tilde{B} = \{(\beta,\beta') \in B \times B : |\beta \cap \beta'| = 1; |V(\beta) \cap V(\beta')| = 2\}$$

and extend the vertex pairs of TTs. Let

$$(\beta,\beta') = \beta \cup \beta' \quad \text{and} \quad V((\beta,\beta')) = V(\beta) \cup V(\beta').$$

Using these definitions, $\breve{V}_S^{(b)}$ can be written as $\lim_{n\to\infty}\sum_{k\in\tilde{B}} U_k$ and it suffices to show that

$$\frac{\lim_{n\to\infty}\mathrm{E}\left\{\sum_{k,l\in\tilde{B}}(U_k - \mathrm{E}\,U_k)(U_l - \mathrm{E}\,U_l)\right\}}{\left(\lim_{n\to\infty}\mathrm{E}\,\sum_{k\in\tilde{B}} U_k\right)^2} \to 0.$$

Note that for $\mathrm{E}(U_k - \mathrm{E}\,U_k)(U_l - \mathrm{E}\,U_l) \neq 0$ we require $V(k) \cap V(l) \neq \varnothing$. Hence, pairs k, l giving non-zero expectation have to comprise at most 5 vertices and are therefore at most of order n^5. □

3.C.3 Lemmas

Lemma 3.2 *Let Υ denote an $n \times n$ random matrix with entries $\Upsilon_{i,j} = Y_{i,j}$ for independent, mean-zero random variables $Y_{i,j}$ that satisfy $\operatorname{E} Y_{i,j}^4 \leq C$ for a finite constant C. Moreover, let M denote a random matrix with $\|M\|_{max} = O_p(n)$ and let D denote a random diagonal matrix with $\|D\|_{max} = O_p(1)$. Then for $A, B \in \{\Upsilon, \Upsilon'\}$*

$$\begin{aligned} (A\,\iota_n)' MB\,\iota_n &= o_p\left(n^2\right), \\ (A\,\iota_n)' DB\,\iota_n &= \operatorname{E}(A\,\iota_n)' DB\,\iota_n + o_p\left(n^2\right), \\ \textit{and} \quad \operatorname{E}(\Upsilon\,\iota_n)' D\Upsilon'\,\iota_n &= o\left(n^2\right). \end{aligned}$$

Proof It suffices to consider the cases $A = B = \Upsilon$ (case 1), and $A = \Upsilon$ and $B = \Upsilon'$ (case 2). Let a_i and b_j denote generic columns of A and B, respectively. Write

$$W^n = \frac{1}{n^2}(A\,\iota_n)' MB\,\iota_n = \frac{1}{n^2}\sum_{i,j}\frac{1}{n}a_i' b_j\,[n\,m_{i,j}] = \frac{1}{n^2}\sum_{i,j}\frac{1}{n}w_{i,j}$$

for $w_{i,j} = \frac{1}{n}\sum_k a_{i,k}b_{j,k}$. By assumption, there is a positive constant C_1 such $-C_1 < n\,m_{i,j} < C_1$, uniformly in i, j. For case 1, $w_{i,j} = n^{-1}\sum_k Y_{k,i}Y_{k,j}$ and $\operatorname{E} w_{i,j} = 0$ for $i \neq j$ and $\operatorname{E} w_{i,j}$ is uniformly bounded otherwise and therefore $\operatorname{E} W^n = o(1)$. Moreover,

$$\operatorname{E} w_{i,j}w_{i',j'} = \frac{1}{n^2}\sum_k \operatorname{E} Y_{k,i}Y_{k,j}Y_{k,i'}Y_{k,j'} + \frac{1}{n^2}\sum_{k \neq k'} \operatorname{E} Y_{k,i}Y_{k,j}Y_{k',i'}Y_{k',j'} = O\left(n^{-1}\right) + 0$$

uniformly over all i, i', j, j' such that $i \neq i'$ or $j \neq j'$ and uniformly bounded otherwise. This implies $\operatorname{E}(W^n)^2 = o(1)$. For case 2 $w_{i,j} = n^{-1}\sum_k Y_{k,i}Y_{j,k}$. Note that $\operatorname{E} w_{i,j} = 0$ if either $i \neq k$ or $j \neq k$ and $\operatorname{E} w_{i,j}$ is uniformly bounded otherwise. Hence $\operatorname{E} W^n = o(1)$. The term $\operatorname{E} w_{i,j}w_{i',j'}$ can be bounded as above. The other statements can be proven in a similar way. □

Appendix 3.D Figures

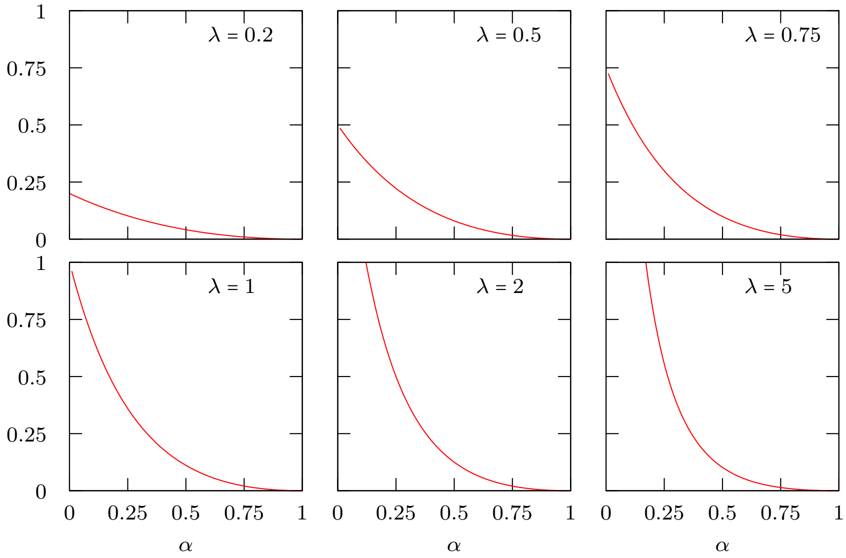


Figure 3.6: The function e from Example 3.1 plotted for various fixed values of λ.

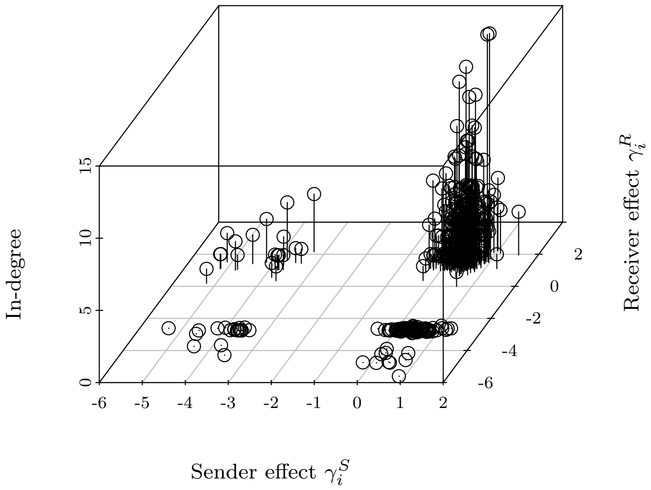


Figure 3.7: Unobserved type vs. observed in-degree for village 60.

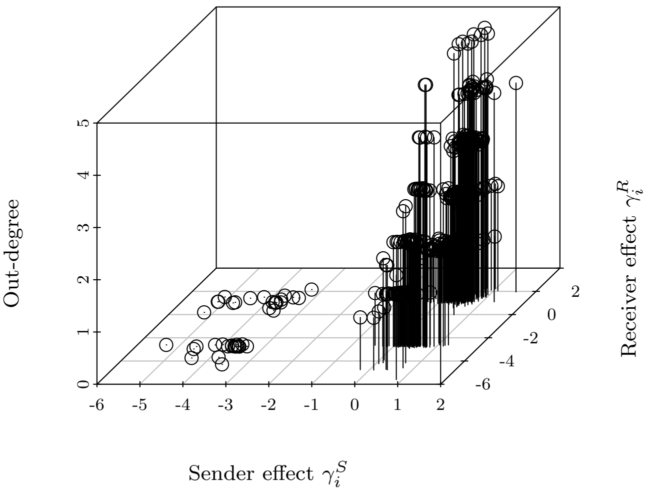


Figure 3.8: Unobserved type vs. observed out-degree for village 60.

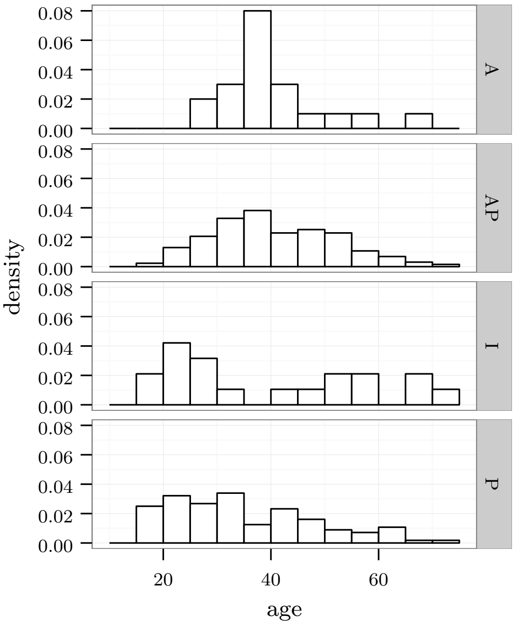


Figure 3.9: Age profiles by cluster for village 60. The unobserved type clusters are: attractor-producers (AP), attractors (A), producers(P) and isolates (I).

Appendix 3.E Tables

Village	n	S_n	$\widehat{\mathrm{E}\,S_n}^A$	$\widehat{\mathrm{E}\,S_n}$	$\widehat{\mathrm{E}\,S_n}^P$	T_n^A	T_n^P
1	203	58	62	23	6	-0.31	37.89
2	203	32	61	20	6	-1.17	12.21
3	345	50	56	15	5	-0.42	51.75
4	256	52	80	18	7	-0.86	27.60
5	164	18	64	11	4	-2.27	12.88
6	110	17	68	12	5	-4.47	9.39
7	172	96	133	39	19	-0.73	20.63
8	109	47	106	33	17	-2.17	7.12
9	247	67	99	23	8	-0.83	38.68
11	142	46	164	30	14	-3.58	9.54
12	195	76	96	23	12	-1.08	18.34
14	150	93	195	53	17	-2.68	19.52
15	212	36	141	28	13	-4.71	6.70
16	178	83	151	46	19	-2.38	17.24
17	200	40	86	28	10	-2.29	13.74
18	284	32	101	21	8	-2.77	14.01
19	243	77	150	41	20	-2.16	13.79
20	159	69	143	42	14	-3.17	17.98
21	210	46	132	26	9	-2.85	17.62
23	280	84	132	26	9	-1.65	41.66
25	304	61	114	25	10	-1.06	30.07
26	149	67	116	31	14	-1.09	18.19
27	174	32	170	24	12	-1.11	6.58
28	395	66	83	25	8	-0.55	41.61
29	303	123	211	49	24	-1.42	21.32
30	170	94	287	34	16	-2.21	24.48
33	219	82	137	36	15	-1.25	19.92
34	181	93	282	33	19	-1.65	18.30
35	216	136	143	47	18	-0.17	32.07
36	293	245	239	92	29	0.11	56.52
37	132	108	114	43	17	-0.14	26.71
38	182	34	134	25	10	-2.66	10.26
39	370	117	173	46	23	-1.26	20.00
40	266	355	267	91	45	1.27	50.65
41	181	272	227	74	37	0.67	36.60
42	206	131	160	54	30	-0.49	18.02
43	227	226	170	55	27	0.95	51.46
44	258	245	163	65	32	1.60	45.06
45	263	66	143	24	11	-1.52	20.84
48	217	107	156	57	26	-1.08	18.01
49	184	79	102	44	25	-0.62	11.47
50	261	259	216	78	48	0.63	33.18
51	309	298	254	108	56	0.69	33.83
52	395	344	329	124	53	0.17	50.43
53	170	183	213	67	35	-0.36	26.43
54	124	64	159	49	27	-1.56	6.95

Continued on next page

Table 3.5 – continued from previous page

Village	n	S_n	$\widehat{\mathrm{E}\,S_n}^A$	$\widehat{\mathrm{E}\,S_n}$	$\widehat{\mathrm{E}\,S_n}^P$	T_n^A	T_n^P
55	279	201	249	71	37	-0.52	29.68
60	413	151	259	71	35	-1.39	20.59
62	242	161	138	52	25	0.52	34.35
64	294	158	155	39	17	0.07	50.28
65	341	344	325	115	57	0.20	38.40
66	189	41	67	19	7	-1.21	23.28
67	231	33	72	19	7	-1.38	15.77
68	164	17	119	21	9	-2.34	4.24
69	220	281	324	132	70	-0.45	21.51
71	298	169	203	55	32	-0.35	24.10
72	238	50	149	25	10	-2.12	17.30
73	217	98	142	47	21	-1.14	20.68
74	193	109	450	45	28	-1.88	14.39
76	269	137	159	41	20	-0.46	34.51
77	172	98	164	52	24	-1.15	17.40

Table 3.5: Estimating and testing predicted transitivity. Estimates for predicted transitivity with bias correction ($\widehat{\mathrm{E}\,S_n}^A$) and without bias correction ($\widehat{\mathrm{E}\,S_n}$). The transitivity estimate for the model without unobserved effects is given by $\widehat{\mathrm{E}\,S_n}^P$. Test statistics for the model with and without unobserved effects are given by T_n^A and T_n^P, respectively.

	A	AP	I	P
age	39	39	39	34
house has own latrine	0.60	0.42	0.79	0.64
no. of rooms	3.65	2.57	3.16	3.29
has savings account	0.30	0.40	0.21	0.27
participates in SHG	0.00	0.12	0.00	0.04
female	0.40	0.54	0.47	0.61
household head	0.45	0.42	0.32	0.27
spouse of household head	0.35	0.42	0.21	0.24
scheduled caste or tribe	0.20	0.32	0.16	0.25
general caste	0.05	0.02	0.00	0.02

Table 3.6: Village 60: means of observed covariates by type cluster (A = attractors, AP = attractor-producers, I = isolates, P = producers).

Variable	Description
age	age of respondent
age2	square of age
female	respondent is female
latrine	respondent lives in a house with an own latrine
obc	respondent's caste is considered an OBC (Other Backward Caste)
general	respondent's caste is considered a General caste
educ Primary	respondent has completed primary education
educ SSLC	respondent has obtained a Secondary Schooling Leaving Certificate
has savings	respondent has at least one savings account
has shg	respondent participates in a SHG (Self Help Group)
is hhhead	respondent is head of her household
is village native	respondent was born in village

Table 3.7: Description of variables approximating productivity (X_i) and popularity (X_j).

References

Aguirregabiria, Victor and Pedro Mira (2007). "Sequential estimation of dynamic discrete games". In: *Econometrica* 75.1, pp. 1–53.

Andersen, Erling (1970). "Asymptotic properties of conditional maximum-likelihood estimators". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 283–301.

Apicella, Coren et al. (2012). "Social networks and cooperation in hunter-gatherers". In: *Nature* 481.7382, pp. 497–501.

Banerjee, Abhijit et al. (2013). "The diffusion of microfinance". In: *Science* 341.6144.

Bearman, Peter, James Moody, and Katherine Stovel (2004). "Chains of affection: The structure of adolescent romantic and sexual networks". In: *American Journal of Sociology* 110.1, pp. 44–91.

Becker, Gary (1973). "A theory of marriage: Part I". In: *The Journal of Political Economy*, pp. 813–846.

Bhamidi, Shankar, Guy Bresler, and Allan Sly. "Mixing time of exponential random graphs". In: *The Annals of Applied Probability.*

Chandrasekhar, Arun and Matthew Jackson (2014). "Tractable and consistent random graph models".

Chandrasekhar, Arun and Randall Lewis (2011). "Econometrics of sampled networks". Working paper.

Charbonneau, Karyne (2014). "Multiple fixed effects in nonlinear panel data models". Working paper.

Davis, James (1970). "Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices". In: *American Sociological Review*, pp. 843–851.

Dhaene, Geert and Koen Jochmans (2010). "Split-panel jackknife estimation of fixed-effect models". Working paper.

Duijn, Marijtje, Tom Snijders, and Bonne Zijlstra (2004). “p2: a random effects model with covariates for directed graphs”. In: *Statistica Neerlandica* 58.2, pp. 234–254.

Erdős, Paul and Alfréd Rényi (1960). “On the evolution of random graphs”. In: *Publications of the Mathematical Institute of the Hungarian Academy of Sciences.*

Fafchamps, Marcel and Flore Gubert (2007). “The formation of risk sharing networks”. In: *Journal of Development Economics* 83.2, pp. 326–350.

Fafchamps, Marcel and Susan Lund (2003). “Risk-sharing networks in rural Philippines”. In: *Journal of Development Economics* 71.2, pp. 261–287.

Fernández-Val, Iván (2009). “Fixed effects estimation of structural parameters and marginal effects in panel probit models”. In: *Journal of Econometrics* 150.1, pp. 71–85.

Fernández-Val, Iván and Martin Weidner (2014). “Individual and time effects in nonlinear panel models with large N, T”. Working paper.

Graham, Bryan (2014). “An empirical model of network formation: detecting homophily when agents are heterogeneous”. Working paper.

Hahn, Jinyong and Guido Kuersteiner (2011). “Bias reduction for dynamic nonlinear panel models with fixed effects”. In: *Econometric Theory* 27.06, pp. 1152–1191.

Hahn, Jinyong and Whitney Newey (2004). “Jackknife and analytical bias reduction for nonlinear panel models”. In: *Econometrica* 72.4, pp. 1295–1319.

Hoff, Peter (2005). “Bilinear mixed-effects models for dyadic data”. In: *Journal of the American Statistical Association* 100.469, pp. 286–295.

Hoff, Peter, Adrian Raftery, and Peter Handcock (2002). “Latent space approaches to social network analysis”. In: *Journal of the american Statistical association* 97.460, pp. 1090–1098.

Holland, Paul and Samual Leinhardt (1970). “A method for detecting structure in sociometric data”. In: *American Journal of Sociology*, pp. 492–513.

— (1976). “Local structure in social networks”. In: *Sociological Methodology* 7, pp. 1–45.

— (1978). “An omnibus test for social structure using triads”. In: *Sociological Methods & Research* 7.2, pp. 227–256.

— (1981). “An exponential family of probability distributions for directed graphs”. In: *Journal of the American Statistical Association* 76.373, pp. 33–50.

Jackson, Matthew (2008). *Social and economic networks.* Princeton University Press.

Jackson, Matthew, Tomas Rodriguez-Barraquer, and Xu Tan (2012). “Social capital and social quilts: Network patterns of favor exchange”. In: *The American Economic Review* 102.5, pp. 1857–1897.

Karlberg, Martin (1997). “Testing transitivity in graphs”. In: *Social Networks* 19.4, pp. 325–343.

— (1999). “Testing transitivity in digraphs”. In: *Sociological Methodology* 29.1, pp. 225–251.

Kim, Min Seong and Yixiao Sun (2013). “Bootstrap and k-step bootstrap bias correction for fixed effects estimators in nonlinear panel models”. Working paper.

Krivitsky, Pavel et al. (2009). “Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models”. In: *Social networks* 31.3, pp. 204–213.

Leung, Michael (2014). "Two-step estimation of network-formation models with incomplete information". Working Paper.

Mayer, Adalbert and Steven Puller (2008). "The old boy (and girl) network: Social network formation on university campuses". In: *Journal of Public Economics* 92.1, pp. 329–347.

McPherson, Miller, Lynn Smith-Lovin, and James Cook (2001). "Birds of a feather: Homophily in social networks". In: *Annual Review of Sociology*, pp. 415–444.

Mele, Angelo (2013). "A structural model of segregation in social networks". Working paper.

Miyauchi, Yuhei (2014). "Structural Estimation of a Pairwise Stable Network with Nonnegative Externality". Working paper.

Neyman, Jerzy and Elizabeth L Scott (1948). "Consistent estimates based on partially consistent observations". In: *Econometrica: Journal of the Econometric Society*, pp. 1–32.

Sheng, Shuyang (2014). "Identification and Estimation of Network Formation Games". Working paper.

Snijders, Tom et al. (2006). "New specifications for exponential random graph models". In: *Sociological Methodology* 36.1, pp. 99–153.

van der Vaart, Aad (2000). *Asymptotic Statistics*. Cambridge University Press.

Wasserman, Stanley and Philippa Pattison (1996). "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p". In: *Psychometrika* 61.3, pp. 401–425.

Watts, Duncan and Steven Strogatz (1998). "Collective dynamics of "small-world" networks". In: *Nature* 393.6684, pp. 440–442.

Bibliography

Abadie, Alberto (2003). "Semiparametric instrumental variable estimation of treatment response models". In: *Journal of Econometrics* 113.2, pp. 231–263.

Aguirregabiria, Victor and Pedro Mira (2007). "Sequential estimation of dynamic discrete games". In: *Econometrica* 75.1, pp. 1–53.

Andersen, Erling (1970). "Asymptotic properties of conditional maximum-likelihood estimators". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 283–301.

Angrist, Joshua, Guido Imbens, and Donald Rubin (1996). "Identification of Causal Effects Using Instrumental Variables". In: *Journal of the American Statistical Association* 91.434, pp. 444–455.

Apicella, Coren et al. (2012). "Social networks and cooperation in hunter-gatherers". In: *Nature* 481.7382, pp. 497–501.

Balke, Alexander and Judea Pearl (1997). "Bounds on treatment effects from studies with imperfect compliance". In: *Journal of the American Statistical Association* 92.439, pp. 1171–1176.

Banerjee, Abhijit et al. (2013). "The diffusion of microfinance". In: *Science* 341.6144.

Bearman, Peter, James Moody, and Katherine Stovel (2004). "Chains of affection: The structure of adolescent romantic and sexual networks". In: *American Journal of Sociology* 110.1, pp. 44–91.

Becker, Gary (1973). "A theory of marriage: Part I". In: *The Journal of Political Economy*, pp. 813–846.

Bhamidi, Shankar, Guy Bresler, and Allan Sly. "Mixing time of exponential random graphs". In: *The Annals of Applied Probability.*

Carneiro, Pedro, James Heckman, and Edward Vytlacil (2011). "Estimating Marginal Returns to Education". In: *American Economic Review* 101.6, pp. 2754–2781.

Carneiro, Pedro and Sokbae Lee (2009). "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality". In: *Journal of Econometrics* 149.2, pp. 191–208.

Chandrasekhar, Arun and Matthew Jackson (2014). "Tractable and consistent random graph models".

Chandrasekhar, Arun and Randall Lewis (2011). "Econometrics of sampled networks". Working paper.

Charbonneau, Karyne (2014). “Multiple fixed effects in nonlinear panel data models”. Working paper.

Chen, Song Xi and Ingrid Van Keilegom (2009). “A goodness-of-fit test for parametric and semi-parametric models in multiresponse regression”. In: *Bernoulli* 15.4, pp. 955–976.

Davis, James (1970). “Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices”. In: *American Sociological Review*, pp. 843–851.

de Jong, Peter (1987). “A central limit theorem for generalized quadratic forms”. In: *Probability Theory and Related Fields* 75.2, pp. 261–277.

Delgado, Miguel A (1993). “Testing the equality of nonparametric regression curves”. In: *Statistics & probability letters* 17.3, pp. 199–204.

Delgado, Miguel A and Wenceslao González Manteiga (2001). “Significance testing in nonparametric regression based on the bootstrap”. In: *Annals of Statistics*, pp. 1469–1507.

Dette, Holger and Natalie Neumeyer (2001). “Nonparametric analysis of covariance”. In: *the Annals of Statistics* 29.5, pp. 1361–1400.

Dhaene, Geert and Koen Jochmans (2010). “Split-panel jackknife estimation of fixed-effect models”. Working paper.

Duijn, Marijtje, Tom Snijders, and Bonne Zijlstra (2004). “p2: a random effects model with covariates for directed graphs”. In: *Statistica Neerlandica* 58.2, pp. 234–254.

Dzemski, Andreas and Florian Sarnetzki (2014). “Overidentification test in a nonparametric treatment model with unobserved heterogeneity”. Working Paper.

Erdős, Paul and Alfréd Rényi (1960). “On the evolution of random graphs”. In: *Publications of the Mathematical Institute of the Hungarian Academy of Sciences.*

Escanciano, Juan Carlos, David Jacho-Chávez, and Arthur Lewbel (2014). “Uniform convergence of weighted sums of non and semiparametric residuals for estimation and testing”. In: *Journal of Econometrics* 178, pp. 426–443.

Escanciano, Juan Carlos and Kyungchul Song (2010). “Testing single-index restrictions with a focus on average derivatives”. In: *Journal of Econometrics* 156.2, pp. 377–391.

Fafchamps, Marcel and Flore Gubert (2007). “The formation of risk sharing networks”. In: *Journal of Development Economics* 83.2, pp. 326–350.

Fafchamps, Marcel and Susan Lund (2003). “Risk-sharing networks in rural Philippines”. In: *Journal of Development Economics* 71.2, pp. 261–287.

Fernández-Val, Iván (2009). “Fixed effects estimation of structural parameters and marginal effects in panel probit models”. In: *Journal of Econometrics* 150.1, pp. 71–85.

Fernández-Val, Iván and Josh Angrist (2013). “ExtrapoLATE-ing: External validity and overidentification in the LATE framework”. In: Tenth World Congress. Advances in Economics and Econometrics: Theory and Applications 3. Econometric Society Monographs.

Fernández-Val, Iván and Martin Weidner (2014). “Individual and time effects in nonlinear panel models with large N, T”. Working paper.

Frölich, Markus (2007). “Nonparametric IV estimation of local average treatment effects with covariates”. In: *Journal of Econometrics* 139.1, pp. 35–75.

Gørgens, Tue (2002). "Nonparametric comparison of regression curves by local linear fitting". In: *Statistics & probability letters* 60.1, pp. 81–89.

Graham, Bryan (2014). "An empirical model of network formation: detecting homophily when agents are heterogeneous". Working paper.

Hahn, Jinyong and Guido Kuersteiner (2011). "Bias reduction for dynamic nonlinear panel models with fixed effects". In: *Econometric Theory* 27.06, pp. 1152–1191.

Hahn, Jinyong and Whitney Newey (2004). "Jackknife and analytical bias reduction for nonlinear panel models". In: *Econometrica* 72.4, pp. 1295–1319.

Hall, Peter and Jeffrey D Hart (1990). "Bootstrap test for difference between means in nonparametric regression". In: *Journal of the American Statistical Association* 85.412, pp. 1039–1049.

Hall, Peter and Joel Horowitz (2012). *A simple bootstrap method for constructing nonparametric confidence bands for functions.* Tech. rep. working paper.

Hansen, Bruce (2008). "Uniform convergence rates for kernel estimation with dependent data". In: *Econometric Theory* 24.03, pp. 726–748.

Hansen, Lars Peter (1982). "Large sample properties of generalized method of moments estimators". In: *Econometrica*, pp. 1029–1054.

Härdle, Wolfgang and Enno Mammen (1993). "Comparing nonparametric versus parametric regression fits". In: *The Annals of Statistics* 21.4, pp. 1926–1947.

Härdle, Wolfgang and James Marron (1985). "Optimal bandwidth selection in nonparametric regression function estimation". In: *The Annals of Statistics*, pp. 1465–1481.

Heckman, James, Daniel Schmierer, and Sergio Urzua (2010). "Testing the correlated random coefficient model". In: *Journal of Econometrics* 158.2, pp. 177–203.

Heckman, James, Sergio Urzua, and Edward Vytlacil (2006). "Understanding instrumental variables in models with essential heterogeneity". In: *The Review of Economics and Statistics* 88.3, pp. 389–432.

Heckman, James and Edward Vytlacil (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation". In: *Econometrica*, pp. 669–738.

Heckman, James et al. (1996). "Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method". In: *Proceedings of the National Academy of Sciences* 93.23, pp. 13416–13420.

— (1998). "Characterizing selection bias using experimental data". In: *Econometrica: Journal of the Econometric Society* 66.5, pp. 1017–1098.

Hoff, Peter (2005). "Bilinear mixed-effects models for dyadic data". In: *Journal of the American Statistical Association* 100.469, pp. 286–295.

Hoff, Peter, Adrian Raftery, and Peter Handcock (2002). "Latent space approaches to social network analysis". In: *Journal of the american Statistical association* 97.460, pp. 1090–1098.

Hoffman, Saul D (1998). "Teenage childbearing is not so bad after all... or is it? A review of the new literature". In: *Family Planning Perspectives* 30.5, pp. 236–243.

Holland, Paul and Samual Leinhardt (1970). "A method for detecting structure in sociometric data". In: *American Journal of Sociology*, pp. 492–513.

Holland, Paul and Samual Leinhardt (1976). “Local structure in social networks”. In: *Sociological Methodology* 7, pp. 1–45.

— (1978). “An omnibus test for social structure using triads”. In: *Sociological Methods & Research* 7.2, pp. 227–256.

— (1981). “An exponential family of probability distributions for directed graphs”. In: *Journal of the American Statistical Association* 76.373, pp. 33–50.

Hotz, V Joseph, Susan Williams McElroy, and Seth G Sanders (2005). “Teenage Childbearing and Its Life Cycle Consequences Exploiting a Natural Experiment”. In: *Journal of Human Resources* 40.3, pp. 683–715.

Hotz, V Joseph, Charles H Mullin, and Seth G Sanders (1997). “Bounding causal effects using data from a contaminated natural experiment: analysing the effects of teenage childbearing”. In: *The Review of Economic Studies* 64.4, pp. 575–603.

Huber, Martin and Giovanni Mellace (2014). “Testing instrument validity for LATE identification based on inequality moment constraints”. In: *Review of Economics and Statistics.*

Ichimura, Hidehiko (1993). “Semiparametric least squares (SLS) and weighted SLS estimation of single-index models”. In: *Journal of Econometrics* 58.1, pp. 71–120.

Imbens, Guido and Joshua Angrist (1994). “Identification and estimation of local average treatment effects”. In: *Econometrica*, pp. 467–475.

Jackson, Matthew (2008). *Social and economic networks.* Princeton University Press.

Jackson, Matthew, Tomas Rodriguez-Barraquer, and Xu Tan (2012). “Social capital and social quilts: Network patterns of favor exchange”. In: *The American Economic Review* 102.5, pp. 1857–1897.

Jackson, Matthew and Asher Wolinsky (1996). “A strategic model of social and economic networks”. In: *Journal of economic theory* 71.1, pp. 44–74.

Jones, Chris, James Marron, and Simon Sheather (1996). “A brief survey of bandwidth selection for density estimation”. In: *Journal of the American Statistical Association* 91.433, pp. 401–407.

Karlberg, Martin (1997). “Testing transitivity in graphs”. In: *Social Networks* 19.4, pp. 325–343.

— (1999). “Testing transitivity in digraphs”. In: *Sociological Methodology* 29.1, pp. 225–251.

Kim, Min Seong and Yixiao Sun (2013). “Bootstrap and k-step bootstrap bias correction for fixed effects estimators in nonlinear panel models”. Working paper.

King, Eileen, Jeffrey D Hart, and Thomas E Wehrly (1991). “Testing the equality of two regression curves using linear smoothers”. In: *Statistics & Probability Letters* 12.3, pp. 239–247.

Kitagawa, Toru (2013). “A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model”. Working Paper.

Klein, Roger W and Richard H Spady (1993). “An efficient semiparametric estimator for binary response models”. In: *Econometrica*, pp. 387–421.

Klepinger, Daniel H, Shelly Lundberg, and Robert D Plotnick (1995). “Adolescent fertility and the educational attainment of young women”. In: *Family planning perspectives*, pp. 23–28.

Kong, Efang, Oliver Linton, and Yingcun Xia (2010). "Uniform bahadur representation for local polynomial estimates of M-regression and its application to the additive model". In: *Econometric Theory* 26.05, pp. 1529–1564.

Krivitsky, Pavel et al. (2009). "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models". In: *Social networks* 31.3, pp. 204–213.

Lee, Ying-Ying (2013). "Partial mean processes with generated regressors: Continuous Treatment Effects and Nonseparable models." Working Paper.

Leung, Michael (2014). "Two-step estimation of network-formation models with incomplete information". Working Paper.

Levine, David I and Gary Painter (2003). "The schooling costs of teenage out-of-wedlock childbearing: analysis with a within-school propensity-score-matching estimator". In: *Review of Economics and Statistics* 85.4, pp. 884–900.

Maistre, Samuel and Valentin Patilea (2014). "Nonparametric model checks for single-index assumptions". Working Paper.

Mammen, Enno (1993). "Bootstrap and wild bootstrap for high dimensional linear models". In: *The Annals of Statistics*, pp. 255–285.

Mammen, Enno, Christoph Rothe, and Melanie Schienle (2012). "Nonparametric regression with nonparametrically generated covariates". In: *The Annals of Statistics* 40.2, pp. 1132–1170.

— (2015). "Semiparametric estimation with generated covariates". In: *Econometric Theory.*

Masry, Elias (1996). "Multivariate local polynomial regression for time series: uniform strong consistency and rates". In: *Journal of Time Series Analysis* 17.6, pp. 571–599.

Mayer, Adalbert and Steven Puller (2008). "The old boy (and girl) network: Social network formation on university campuses". In: *Journal of Public Economics* 92.1, pp. 329–347.

McPherson, Miller, Lynn Smith-Lovin, and James Cook (2001). "Birds of a feather: Homophily in social networks". In: *Annual Review of Sociology*, pp. 415–444.

Mele, Angelo (2013). "A structural model of segregation in social networks". Working paper.

Miller, Amalia R (2011). "The effects of motherhood timing on career path". In: *Journal of Population Economics* 24.3, pp. 1071–1100.

Miyauchi, Yuhei (2014). "Structural Estimation of a Pairwise Stable Network with Nonnegative Externality". Working paper.

Neumeyer, Natalie and Holger Dette (2003). "Nonparametric comparison of regression curves: an empirical process approach". In: *The Annals of Statistics* 31.3, pp. 880–920.

Neyman, Jerzy and Elizabeth L Scott (1948). "Consistent estimates based on partially consistent observations". In: *Econometrica: Journal of the Econometric Society*, pp. 1–32.

Nolan, Deborah and David Pollard (1987). "U-processes: Rates of Convergence". In: *The Annals of Statistics*, pp. 780–799.

Pollard, David (1984). *Convergence of stochastic processes.* Springer.

Reinhold, Steffen (2007). “Essays in demographic Economics”. PhD thesis. John Hopkins University.

Ribar, David C (1994). “Teenage fertility and high school completion”. In: *The Review of Economics and Statistics*, pp. 413–424.

Ruppert, David and Matthew P Wand (1994). “Multivariate locally weighted least squares regression”. In: *The Annals of Statistics*, pp. 1346–1370.

Sargan, John (1958). “The estimation of economic relationships using instrumental variables”. In: *Econometrica: Journal of the Econometric Society*, pp. 393–415.

Sheng, Shuyang (2014). “Identification and Estimation of Network Formation Games”. Working paper.

Sherman, Robert P (1994). “Maximal inequalities for degenerate U-processes with applications to optimization estimators”. In: *The Annals of Statistics*, pp. 439–459.

Snijders, Tom et al. (2006). “New specifications for exponential random graph models”. In: *Sociological Methodology* 36.1, pp. 99–153.

Stute, Winfried and Li-Xing Zhu (2005). “Nonparametric checks for single-index models”. In: *Annals of Statistics*, pp. 1048–1083.

van de Geer, Sara (2000). *Empirical Processes in M-estimation.* Vol. 6. Cambridge University Press.

van der Vaart, Aad (2000). *Asymptotic Statistics.* Cambridge University Press.

van der Vaart, Aad and Jon Wellner (1996). *Weak Convergence and Empirical Processes.* Springer.

Vytlacil, Edward (2002). “Independence, monotonicity, and latent index models: An equivalence result”. In: *Econometrica* 70.1, pp. 331–341.

Vytlacil, Edward and Nese Yildiz (2007). “Dummy endogenous variables in weakly separable models”. In: *Econometrica* 75.3, pp. 757–779.

Wasserman, Stanley and Philippa Pattison (1996). “Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p”. In: *Psychometrika* 61.3, pp. 401–425.

Watts, Duncan and Steven Strogatz (1998). “Collective dynamics of “small-world” networks”. In: *Nature* 393.6684, pp. 440–442.

Xia, Yingcun et al. (2004). “A goodness-of-fit test for single-index models”. In: *Statistica Sinica* 14.1, pp. 1–28.